# Toward Cognitive and Immersive Systems: Experiments in a Cognitive Microworld

**Matthew Peveler** PEVELM@RPI.EDU
**Naveen Sundar Govindarajulu** NAVEEN.SUNDAR.G@GMAIL.COM
**Selmer Bringsjord** SELMER.BRINGSJORD@GMAIL.COM
**Atriya Sen** ATRIYA@ATRIYASEN.COM
Rensselaer AI & Reasoning Lab, Department of Computer Science,
Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, USA

**Biplav Srivastava** BIPLAVS@US.IBM.COM
**Kartik Talamadupula** KRTALAMAD@US.IBM.COM
**Hui Su** HUISUIBMRES@US.IBM.COM
IBM TJ Watson Research Center, Yorktown Heights, NY 10598, USA

## Abstract

As computational power has continued to increase, and sensors have become more accurate, the corresponding advent of systems that are at once cognitive and immersive has arrived. These *cognitive and immersive systems* (CAISs) fall squarely into the intersection of AI with HCI/HRI: such systems interact with and assist the human agents that enter them, in no small part because such systems are infused with AI able to understand and reason about these humans and their knowledge, beliefs, goals, communications, plans, etc. We herein explain our approach to engineering CAISs. We emphasize the capacity of a CAIS to develop and reason over a “theory of the mind” of its human partners. This capacity entails that the AI in question has a sophisticated model of the beliefs, knowledge, goals, desires, emotions, etc. of these humans. To accomplish this engineering, a formal framework of very high expressivity is needed. In our case, this framework is a *cognitive event calculus*, a particular kind of quantified multi-operator modal logic, and a matching high-expressivity automated reasoner and planner. To explain, advance, and to a degree validate our approach, we show that a calculus of this type satisfies a set of formal requirements, and can enable a CAIS to understand a psychologically tricky scenario couched in what we call the *cognitive polysolid framework* (CPF). We also formally show that a room that satisfies these requirements can have a useful property we term *expectation of usefulness*. CPF, a sub-class of *cognitive microworlds*, includes machinery able to represent and plan over not merely blocks and actions (such as seen in the primitive “blocks worlds” of old), but also over agents and their mental attitudes about both other agents and inanimate objects.

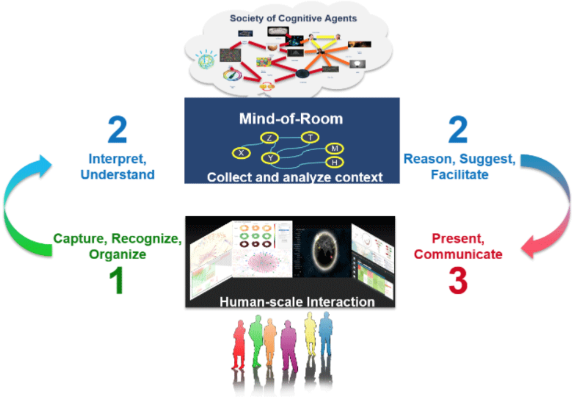

Figure 1: The flow of information through the areas of a CAIS.

## 1. Introduction

In contemporary AI research devoted to decision support, the challenge is often taken to be that of providing AI support to a single human. However, much human problem-solving is fundamentally social, in that a group of people must work together to solve a problem, and must rely upon machine intelligence that is itself highly diverse. Examples of such activities include: hiring a person into a university or company, tackling an emergency crisis like a water pipeline break, planning an intricate medical operation, deciding on companies for merging or acquisition, etc. Motivated by such challenges, we are interested in how an artificial agent — embedded in a social-collaboration environment like an immersive room — can, on the spot, help a group of human participants.

We first introduce the notion of a *cognitive and immersive system* (CAIS), which is comprised of three sub-areas linked to each other in a cyclical flow, shown in Figure 1. The first area is responsible for perception and sensing within the environment housing the human agents (specifically for us, within the room). Percepts come courtesy of a range of sensors; for example, microphones and kinects. The second area covers interpreting, understanding, and acting upon perceived data through reasoning, planning, learning, and NLP. The third area is devoted to displaying both percepts, and the results of processing thereof, in rich multi-modal ways. The particular CAIS we have so far used for our investigations additionally has access to a variety of external machines and services that can be called upon to process requests, queries, and tasks, and to incorporate the results of analysis of additional information from these ines into further decision-making. An important part of our particular CAIS is that there are some number of overseeing AIs (agents) operating at the system level that can make use of any part of the CAIS to assist and aid the humans and other AIs that are operating within the room. Thus this architecture is neither fully centralized, nor fully distributed, but aims to combine the strengths of both.

As an initial test of our CAIS implementation, we examine two scenarios wherein participants have an imbalance of knowledge/beliefs that might influence their actions. It is common in group discussions that not all participants are aware of everything said in the discussion, whether because they missed something being said or misconstrue something. However, they may still try to act on their beliefs that do not (or no longer) properly reflect reality, or may know of their ignorance and wish to remedy it. A CAIS should be able to offer help to participants in these cases, either by alerting participants if they are acting on beliefs that are no longer relevant, or by giving a brief summary of the things that had happened while they were out of the discussion. To do this, a CAIS must be able to model the theory of mind (Premack & Woodruff, 1978; Frith & Frith, 2005; Arkoudas & Bringsjord, 2009a) of the participants and track its state through time. On the basis of its understanding of this modeling, a CAIS must be able to step in as necessary to offer corrective advice, along with a justification for it. Most importantly, the particular capacities we have just enumerated as desiderata for a CAIS must flow from underlying formal requirements that rigorously capture the general desiderata in question.

To accomplish the above tasks, we first present informal requirements that differentiate a CAIS from other intelligent agents. The requirements are in terms of the cognitive states and common knowledge of the agents within a CAIS. As far as we know, this is the first such characterization of what separates an intelligent room from an intelligent agent. We then cast the informal requirements in a formal logic and show that these requirements lead to a useful property (**Property 1**). We then briefly present a framework for a domain of problems that can be used to test a CAIS. From there, we define two tasks with that framework, and then show that by **Property 1**, the system (with the relevant information) can solve the tasks. Finally, after a sustained discussion of related, prior research in AI in which we make clear the unique power of our formalisms and technology, we discuss promising future lines of work.

### 1.1 Cognitive Immersive Room

In development of a CAIS, much work has gone into the creation of a cognitive immersive room architecture (Divekar et al., 2018) which allows for research into augmenting human group collaboration and decision-making with cognitive artificial agents. To start, we build on the prior work in the space of intelligent rooms, primarily that carried out in Brooks (1997).

At the core of our room, we have an array of microphones that hear what people say, which utilizes a transcription service to translate the speech to text. Given the text, we then check for the presence of a "trigger word" to indicate if a user is talking to the room or not. If the "trigger word" is detected, the text is then further analyzed (utilizing the IBM Watson Assistant[1]) to extract an intent for the text as well as any keywords in the text given the command. The intent and keywords are then fed into the executor which in turn drives the room, and can call external services as necessary and output content to users via connected displays or speakers. Additionally, the executor maintains the state of the room as well as definition of the conversation tree that is available to the participants at any given time. Each of these components are implemented and run separately and communication between them is handled via passing JSON objects using the message queue server RabbitMQ[2].

In addition to these, there is a host of external web services made available to any component through the standard GET/POST HTTP request headers. An example of such a web service is the "name-resolver-general," which, given a name, will return a list of probable matches for that

[1] https://www.ibm.com/watson/services/conversation/

[2] https://www.rabbitmq.com

keyword, which allows us to handle misspellings introduced in the keyword by the speech-to-text translation. Each of these external web services are registered in our "service-registry," which the executor references using its name to get the IP and port of these services to be able to call them. The registry also acts as a monitor for each service to determine if the service is available or not. Finally, we utilize the in-memory data store of Redis[3] to store information about our agents during a particular usage or session of the room and then use the PostgreSQL[4] database to store information on a longer-term basis. The displays for the room are three projectors on which we run Electron[5], which allows for building desktop GUIs powered by web technologies, including HTML, CSS, and JS. In Electron, the executor can open various web pages and sites within the GUI, showing both internal (such as transcript or command log) and external content (such as Google maps or Youtube).

### 1.2 Recognition of the Need for Theory-of-Mind-Level AI

Independent of the specific formalisms and corresponding technology that we soon bring to bear herein, we first point out that the need to model the mental states of humans in order to engineer certain AI systems that understand and interact well with these humans has been recognized. For example, work in human-robot teaming has focused on the use of automated planning techniques that take human goals and (mental) states into account. In addition, work on human-aware task planning for mobile robots (Cirillo et al., 2009) has used *predicted* plans for the humans to guide the automated system's own planning. This direction was made more explicit in work on coördinating the goals and plans of humans and robots (Talamadupula et al., 2014; Chakraborti et al., 2015); here, a subset of the humans' mental states relevant to the autonomous system's planning problems was explicitly represented and reasoned with. Very recent work has focused on adapting these previous ideas and techniques to proactive decision-making (Sengupta et al., 2017; Kim & Shah, 2017) and smart-room environments (Chakraborti et al., 2017). Pearce et al. (2014) note the importance of what they dub "social planning," which includes an agent's seeking a goal via the modification of the mental states of others. For a final example, in (Langley et al., 2016), PUG, a system that uses mental simulation to plan for symbolic goals that have numeric utilities, is presented.[6]

On the other hand, while these papers confirm the recognition to which we refer, and while they feature some level of formalization of human-level mental states, they seem to us to lack the necessary technical formal and computational machinery needed to mechanize a full human-level theory of mind — let alone such a theory of mind *and* the requirements of a truly smart room/CAIS we set below.[7] We now turn to the presentation of the requisite formal and computational machinery.

[3] https://redis.io

[4] https://www.postgresql.org

[5] https://electronjs.org

[6] Our latter two examples here are briefly returned to later in "Prior/Related Work and Novelty."

[7] Note that in the present paper there is a limit to the theory-of-mind-modeling "power" we insist an overseeing AI have. E.g., we don't require that an AI overseeing an environment populated with humans have so-called *phenomenal consciousness*, a form of "what's it's like to" consciousness characterized e.g. by Block (1995), and claimed by Bringsjord (2007) to be impossible for a mere machine to possess. In sharp contrast with phenomenal consciousness, *cognitive consciousness* consists only in the logico-mathematical *structure* of human-level (and, indeed, above) cognition, instantiated through time. While cognitive consciousness can be characterized axiomatically with help from

## 2. The Deontic Cognitive Event Calculus

To capture the room in a formal way, we employ the **deontic cognitive event calculus** ($\mathcal{DCEC}$).[8] While the full syntax and inference schemata are outside the scope of this paper, we give a brief overview.[9] $\mathcal{DCEC}$ is a multi-sorted quantified modal logic with a well-defined syntax and proof calculus. $\mathcal{DCEC}$ subsumes the event calculus (Mueller, 2014), a first-order calculus used for modeling events and actions and their effects upon the world. The proof calculus of $\mathcal{DCEC}$ is based on natural deduction (Gentzen, 1964) and includes all the introduction and elimination rules for first-order logic, as well as inference schemata for the modal operators and related structures. $\mathcal{DCEC}$ is a sorted system and includes the following built-in sorts:

| Sort | Description |
|---|---|
| **Agent** | Human and non-human actors. |
| **Time** | Time points. E.g., $t_i$, $birthday(son(jack))$ etc. |
| **Event** | Used for events in the domain. |
| **ActionType** | Abstract actions instantiated by agents. |
| **Action** | Events that occur as actions by agents |
| **Fluent** | Representing states of the world. |

The intensional operators and necessary inference schemata for this paper are shown below below. The operator $\mathbf{B}(a, t, \phi)$ represents that agent $a$ at time $t$ believes $\phi$. The operator $\mathbf{K}(a, t, \phi)$ represents that agent $a$ at time $t$ knows $\phi$.[10] The operator $\mathbf{D}(a, t, \phi)$ represents that agent $a$ at time $t$ desired $\phi$. The operator $\mathbf{C}(t, \phi)$ represents that at time $t$, $\phi$ is common knowledge, which from the inference schemata defined below we see means subsequently that all agents know $\phi$. The operator $\mathbf{S}(a, b, t, \phi)$ represents that agent $a$ told agent $b$ $\phi$ at time $t$. Alternatively, it can be used as $\mathbf{S}(a, t, \phi)$, which represents that agent $a$ at time $t$ said $\phi$ (and everyone hears it). We also have the operator $\mathbf{P}(a, t, \phi)$, which represents that agent $a$ at time $t$ perceived $\phi$ (giving us also that agent $a$ knows $\phi$ at time $t$).

For our current purposes, the main inference schemata needed include $I_{\mathbf{K}}$ and $I_{\mathbf{B}}$, which state that knowledge and belief are closed under the inference system of $\mathcal{DCEC}$. We also have inference schemata that let us go from perception to knowledge ($I_1$), knowledge to belief ($I_2$), common knowledge to knowledge ($I_3$), and from knowledge to propositions that hold ($I_4$). Later below, we also use *derived inference schemata* for converting perceptions to knowledge, knowledge to belief, common knowledge to belief etc., labeled as $D_{[\mathbf{P}\rightsquigarrow\mathbf{K}]}$, $D_{[\mathbf{K}\rightsquigarrow\mathbf{B}]}$, and $D_{[\mathbf{C}\rightsquigarrow\mathbf{B}]}$ respectively (Arkoudas & Bringsjord, 2008).

the formal languages we introduce below for cognitive calculi (Bringsjord et al., 2018), in the present paper we do not require the AI overseeing CAISs to have even cognitive consciousness, and we specifically do not require, at this early point in our work on CAISs, cognitive *self*-consciousness, despite the fact that the latter is something that has been significantly mechanized and implemented Bringsjord et al. (2015); Bringsjord (2010).

[8] We do not use the deontic components in $\mathcal{DCEC}$ in this paper.

[9] For a more in-depth primer on the $\mathcal{DCEC}$, see the appendix in Govindarajulu & Bringsjord (2017a).

[10] Note that knowing is not decomposable or inferable from belief.

**Syntax**

$$
\begin{aligned}
S ::= \;& \textsf{Object} \mid \textsf{Agent} \mid \textsf{Self} \sqsubset \textsf{Agent} \mid \textsf{ActionType} \\
& \textsf{Action} \sqsubseteq \textsf{Event} \mid \textsf{Moment} \mid \textsf{Formula} \mid \textsf{Fluent} \mid \textsf{Numeric} \\
f ::= \;& \mathit{action} : \textsf{Agent} \times \textsf{ActionType} \rightarrow \textsf{Action} \\
& \mathit{initially} : \textsf{Fluent} \rightarrow \textsf{Formula} \\
& \mathit{holds} : \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Formula} \\
& \mathit{happens} : \textsf{Event} \times \textsf{Moment} \rightarrow \textsf{Formula} \\
& \mathit{clipped} : \textsf{Moment} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Formula} \\
& \mathit{initiates} : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Formula} \\
& \mathit{terminates} : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Formula} \\
& \mathit{prior} : \textsf{Moment} \times \textsf{Moment} \rightarrow \textsf{Formula} \\
t ::= \;& x : S \mid c : S \mid f(t_1, \ldots, t_n) \\
& t : \textsf{Formula} \mid \neg\phi \mid \phi \land \psi \mid \phi \lor \psi \mid \\
\phi ::= \;& \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \\
& \mathbf{S}(a,t,\phi) \mid \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\mathit{holds}(f,t')) \mid \mathbf{I}(a,t,\phi)
\end{aligned}
$$

**Inference Schemata** (fragment)

$$\frac{\mathbf{K}(a,t_1,\Gamma),\ \ \Gamma \vdash \phi,\ \ t_1 \leq t_2}{\mathbf{K}(a,t_2,\phi)}\ [I_{\mathbf{K}}]$$

$$\frac{\mathbf{B}(a,t_1,\Gamma),\ \ \Gamma \vdash \phi,\ \ t_1 \leq t_2}{\mathbf{B}(a,t_2,\phi)}\ [I_{\mathbf{B}}]$$

$$\frac{}{\mathbf{C}(t,\mathbf{P}(a,t,\phi) \rightarrow \mathbf{K}(a,t,\phi))}\ [I_1]$$

$$\frac{}{\mathbf{C}(t,\mathbf{K}(a,t,\phi) \rightarrow \mathbf{B}(a,t,\phi))}\ [I_2]$$

$$\frac{\mathbf{C}(t,\phi)\ t \leq t_1 \ldots t \leq t_n}{\mathbf{K}(a_1,t_1,\ldots \mathbf{K}(a_n,t_n,\phi)\ldots)}\ [I_3]$$

$$\frac{\mathbf{K}(a,t,\phi)}{\phi}\ [I_4]$$

$$\vdots \qquad\qquad \vdots$$

## 2.1 Non-modal Systems are not Enough

Note that first-order logic is an **extensional** system; modal logics are **intensional** systems. $\mathcal{DCEC}$ is **intensional** in the sense that it has intensional operators.[11] Formal systems that are intensional are crucial for modeling theory-of-mind reasoning. One simple reason is that using first-order logic leads to unsound inferences as shown below, in which we have an agent $r$ that knows the manager of a team is the most responsible person in the team. Agent $r$ does not know that $Moe$ is the manager of the team, but it's true that $Moe$ is the manager. If the knowledge operator **K** is a simple first-order predicate, we get the proof shown below, which produces a contradiction (that $r$ knows that $Moe$ is the manager) from true premises. This unsoundness persists even with more robust representation schemes in extensional logics (Bringsjord & Govindarajulu, 2012).

$\boxed{1}$ $\mathbf{K}\left(r,\ \textsf{Manager}\left(\mathit{team}, \mathit{mostResponsible}\left(\mathit{team}\right)\right)\right)$ ; given

$\boxed{2}$ $\neg\mathbf{K}\left(r, \textsf{Manager}\left(\mathit{team}, \mathit{Moe}\right)\right)$ ; given

$\boxed{3}$ $\mathit{Moe} = \mathit{mostResponsible}\left(\mathit{team}\right)$ ; given

$\boxed{4}$ $\mathbf{K}\left(r, \textsf{Manager}\left(\mathit{team}, \mathit{Moe}\right)\right)$ ; first-order inference from $\boxed{3}$ and $\boxed{1}$

$\boxed{5}$ $\perp$ ; first-order inference from $\boxed{4}$ and $\boxed{2}$

## 2.2 Reasoner (Theorem Prover)

To handle reasoning within $\mathcal{DCEC}$ we utilize a quantified modal logic theorem prover, ShadowProver.[12] The prover works by utilizing a technique called **shadowing** to achieve speed without

[11] Please note that there is a vast difference between intension and intention.

[12] The prover is available in both Java and Common Lisp and can be obtained at: `https://github.com/naveensundarg/prover`. The underlying first-order prover is SNARK, available at: `http://www.ai.sri.com/~stickel/snark.html`.

sacrificing consistency in the system. Describing the details of the reasoner are beyond the scope here. See (Govindarajulu & Bringsjord, 2017a,b) for more details.

### 2.3 Planner

Planning for the room is handled by Spectra, a planner based on an *extension* of STRIPS-style planning.[13] In this planning formalism, arbitrary formulae of $\mathcal{DCEC}$ are allowed in states, actions, and goals. For instance, valid states and goals can include: *"No three blocks on the table should be of the same color."* and *"Jack believes that Jill believes there is one block on the table."*

## 3. What is a CAIS? Informal Requirements

Note that a CAIS can be considered an intelligent room that specifically requires intelligence of a cognitive sort; that is, it is not sufficient that the room be intelligent about, for example, search queries over a domain $D$; the room should also be intelligent about cognitive states of agents in the room and their cognitive states and attitudes toward $D$.

Despite there being a significant amount of work done in building intelligent environments (of varying levels of intelligence; Coen et al., 1998; Brooks, 1997; Chan et al., 2008), there is no formalization of what constitutes an intelligent room and what separates it from an intelligent agent. Though (Coen et al., 1998) briefly differentiates an intelligent room from ubiquitous computing based on the non-ubiquity of sensors in the former, there is not any formal or rigorous discussion of what separates an intelligent room from a mobile robot that roams around the room with an array of sensors. We offer below a sketch of informal requirements that an immersive room should aim for. Then we instantiate these requirements using $\mathcal{DCEC}$.

The requirements in question are cognitive in nature and exceed intelligent rooms with sensors that can answer queries over simple extensional data (e.g. a room that can answer financial queries such as *"Show me the number of companies with revenue over X?"*). At a high-level, we require that the two conditions below hold:

**Informal Requirements**

$\mathcal{C}$ *Cognitive* A CAIS system should be able to help agents with cognitive tasks and goals. For instance, a system that simply aids in querying a domain $D$ is not cognitive in nature; a system that aids an agent in convincing another agent that some state-of-affairs holds in $D$ is considered cognitive.

$\mathcal{I}$ *Immersive/Non-localized* There should be some attribute or property of a CAIS that is non-localized and distinguished from agents in the room. Moreover, this property should be **common knowledge**. (Note: this is not easily achievable with a physical robot, and this condition differentiates a CAIS system from a cognitive agent.)[a]

[a] This condition may not strictly be realizable, but the goal is to at a minimum build systems that approach this ideal condition.

[13] "STRIPS-style" ascribed to a planner $P$ means that some of the prominent properties that a STRIPS planner has, $P$ has. For illustration by analogy, a theorem prover in the style of some famous first-order one (e.g. Otter) could be capable of reasoning with second-order formulae. A programming language in the logic-programming style could actually fail to be based on first-order logic or a fragment thereof. It is important to note that Spectra's planning is based on an *extension* of STRIPS-style planning.

## 4. Formal Requirements for a CAIS

Now we translate the informal requirements presented above for a CAIS $\gamma$. The CAIS we present acts as an arbiter when goals of agents conflict, and acts to rectify false or missing beliefs. Assume that enclosed within $\gamma$'s space at time $t$ are agents $A(t) = \{a_1, \ldots, a_n\}$. (Note: $\gamma$ is also an agent but is not included in $A(t)$.) Time is assumed to be discrete, as in the discrete event calculus presented in Mueller (2014). There is a background set of axioms and propositions $\Gamma(t)$ that is operational at time $t$. We have a fluent *vicinity* that tells us whether an agent is in the vicinity of a fluent, event, or another agent. Only events and fluents in the vicinity of an agent can be observed by the agent.

$$\mathit{vicinity} : \textsf{Agent} \times \textsf{Fluent} \cup \textsf{Event} \cup \textsf{Agent} \rightarrow \textsf{Fluent}$$

For the cognitive condition $\mathcal{C}$ above, we have the following concrete requirements that we implement in our system. Later, we give examples of these requirements in action.

**Formal Requirements for $\mathcal{C}$**

Assume $\Gamma \vdash t < t + \Delta$

$\mathbf{C}_1^f$ : It is common knowledge that, if an agent $x$ has a false belief, $\gamma$ informs the agent of the belief:[a]

$$\mathbf{C}\left(t, \begin{bmatrix} \mathbf{B}(\gamma, t, \phi) \land \mathbf{B}\big(\gamma, t, \mathbf{B}(x, t, \neg\phi)\big) \\ \rightarrow \\ \mathbf{S}(\gamma, x, t + \Delta, \phi) \end{bmatrix}\right)$$

$\mathbf{C}_2^f$ : It is common knowledge that, if an agent $x$ has a missing belief, $\gamma$ informs the agent of the belief:

$$\mathbf{C}\left(t, \begin{bmatrix} \mathbf{B}(\gamma, t, \phi) \land \mathbf{B}\big(\gamma, t, \neg\mathbf{B}(x, t, \phi)\big) \\ \rightarrow \\ \mathbf{S}(\gamma, x, t + \Delta, \phi) \end{bmatrix}\right)$$

[a] Please note that inference in $\mathcal{DCEC}$ is non-monotonic as it includes the event calculus, which is non-monotonic. If an agent $a$ believes $\phi$ based on prior information, adding new information can cause the agent to not believe $\phi$.

For the immersive condition $\mathcal{I}$ above, we have the following conditions:

**Formal Requirements for $\mathcal{I}$**

$\mathbf{I}_1^f$ : It is common knowledge that at any point in time $t$ an agent $x$, different from the CAIS system $\gamma$, can observe events or conditions (fluents) only in its vicinity.

$$\mathbf{C}\Big(\forall x,t,f : (x \neq \gamma) \rightarrow \Big[\mathbf{P}\big(x,t,\mathit{holds}(f,t)\big) \rightarrow \mathit{holds}(\mathit{vicinity}(x,f),t)\Big]\Big)$$

$$\mathbf{C}\Big(\forall x,t,e : (x \neq \gamma) \rightarrow \Big[\mathbf{P}\big(x,t,\mathit{happens}(e,t)\big) \rightarrow \mathit{holds}(\mathit{vicinity}(x,e),t)\Big]\Big)$$

$\mathbf{I}_2^f$ : It is common knowledge that actions performed by the agent are in its vicinity.

$$\mathbf{C}\Big(\forall x,\alpha : \mathit{holds}\big(\mathit{vicinity}(x,\mathit{action}(a,\alpha)),t\big)\Big)$$

$\mathbf{I}_3^f$ : It is common knowledge that all events and fluents are perceived by $\gamma$. This is represented by the four conditions below:

$$(i) \quad \mathbf{C}\Big(\forall t,f : \Big[\mathit{holds}(f,t) \leftrightarrow \mathbf{P}\big(\gamma,t,\mathit{holds}(f,t)\big)\Big]\Big)$$

$$(ii) \quad \mathbf{C}\Big(\forall t,e : \Big[\mathit{happens}(e,t) \leftrightarrow \mathbf{P}\big(\gamma,t,\mathit{happens}(e,t)\big)\Big]\Big)$$

$$(iii) \quad \mathbf{C}\Big(\forall t,f : \Big[\neg\mathit{holds}(f,t) \leftrightarrow \mathbf{P}\big(\gamma,t,\neg\mathit{holds}(f,t)\big)\Big]\Big)$$

$$(iv) \quad \mathbf{C}\Big(\forall t,e : \Big[\neg\mathit{happens}(e,t) \leftrightarrow \mathbf{P}\big(\gamma,t,\neg\mathit{happens}(e,t)\big)\Big]\Big)$$

## 5. A Foundational Property of CAIS

One of the benefits of having a properly designed CAIS is that humans inside it can rely upon the CAIS to help other agents with relevant missing or false information $\phi$, as opposed to the case with a mobile robot. If we had a localized mobile robot instead of a CAIS, the human would need to decide whether the robot has the required information $\phi$ and needs to believe that the robot believes that the other human is missing the relevant information $\phi$. If a CAIS system satisfies the above condition, we can derive the following foundationally important (object-level in $\mathcal{DCEC}$) property that states this in a formal manner.

**Property 1: Expectation of Usefulness**

If the above properties hold, then an agent $a$ that perceives that another agent $b$ is not aware of an event happening, believes that CAIS $\gamma$ will inform $b$ (Assume: $\Gamma \vdash t < t + \Delta$):

$$\begin{gathered}\mathbf{P}\big(a,t,\mathit{happens}(e,t)\big) \wedge \mathbf{P}\Big(a,t,\neg\mathit{holds}\big(\mathit{vicinity}(b,e),t\big)\Big) \\ \rightarrow \\ \mathbf{B}\Big(a,t,\mathbf{S}\big(\gamma,b,t+\Delta,\mathit{happens}(e,t)\big)\Big)\end{gathered}$$

**Proof Sketch**: Colors used for readability.

$$D_{[\mathbf{K}\rightsquigarrow\mathbf{B}]}\ \dfrac{D_{[\mathbf{P}\rightsquigarrow\mathbf{K}]}\ \dfrac{\mathbf{P}(a,t,\mathit{happens}(e,t))}{\mathbf{K}(a,t,\mathit{happens}(e,t))}}{\mathbf{B}(a,t,\mathit{happens}(e,t)) \equiv \boxed{\phi_1}} \qquad D_{[\mathbf{K}\rightsquigarrow\mathbf{B}]}\ \dfrac{D_{[\mathbf{P}\rightsquigarrow\mathbf{K}]}\ \dfrac{\mathbf{P}(a,t,\neg \mathit{holds}\big(\mathit{vicinity}(b,e),t\big))}{\mathbf{K}\Big(a,t,\neg \mathit{holds}\big(\mathit{vicinity}(b,e),t\big)\Big)}}{\mathbf{B}\Big(a,t,\neg \mathit{holds}\big(\mathit{vicinity}(b,e),t\big)\Big) \equiv \boxed{\phi_2}}$$

Using $\mathbf{I}_3^f(ii)$ the CAIS observes all events that happen in its enclosure:

$$I_{\mathbf{B}}\ \dfrac{I_{\mathbf{B}}\ \dfrac{D_{[\mathbf{C}\rightsquigarrow\mathbf{B}]}\ \dfrac{\mathbf{I}_3^f(ii) \equiv \mathbf{C}\Bigg(\forall t,e : \Big[\mathit{happens}(e,t) \leftrightarrow \mathbf{P}\big(\gamma,t,\mathit{happens}(e,t)\big)\Big]\Bigg)}{\mathbf{B}\ \Big(a,t,\forall t,e : \Big[\mathit{happens}(e,t) \leftrightarrow \mathbf{P}\big(\gamma,t,\mathit{happens}(f,t)\big)\Big]\Big)} \qquad \boxed{\phi_1}}{\mathbf{B}\Big(a,t,\mathbf{P}\big(\gamma,t,\mathit{happens}(e,t)\big)\Big)}}{\mathbf{B}\big(a,t,\mathbf{B}(\gamma,t,\mathit{happens}(e,t))\big) \equiv \boxed{\psi_1}}$$

Similarly, using $\mathbf{I}_3^f(iii)$:

$$I_{\mathbf{B}}\ \dfrac{I_{\mathbf{B}}\ \dfrac{D_{[\mathbf{C}\rightsquigarrow\mathbf{B}]}\ \dfrac{\mathbf{I}_3^f(iii) \equiv \mathbf{C}\Bigg(\forall t,e : \Big[\neg \mathit{holds}(e,t) \leftrightarrow \mathbf{P}\big(\gamma,t,\neg \mathit{holds}(e,t)\big)\Big]\Bigg)}{\mathbf{B}\ \Big(a,t,\forall t,e\Big[\neg \mathit{holds}(e,t) \leftrightarrow \mathbf{P}\big(\gamma,t,\neg \mathit{holds}(e,t)\big)\Big]\Big)} \qquad \boxed{\phi_2}}{\mathbf{B}(a,t,\mathbf{P}(\gamma,t,\neg \mathit{holds}\big(\mathit{vicinity}(b,e),t\big)))}}{\mathbf{B}(a,t,\mathbf{B}\big(\gamma,t,\neg \mathit{holds}\big(\mathit{vicinity}(b,e),t\big)\big) \equiv \boxed{\phi_3}}$$

From $\mathbf{I}_1^f$ (and from $\mathbf{B}\big(a,t,\mathbf{B}(\gamma,t,b \neq \gamma)\big)$):

$$I_{\mathbf{B}}\ \dfrac{I_{\mathbf{B}}\ \dfrac{I_{\mathbf{B}}\ \dfrac{D_{[\mathbf{C}\rightsquigarrow\mathbf{B}]}\ \dfrac{\mathbf{C}\Big(\forall x,t,e : (x \neq \gamma) \rightarrow \big[\mathbf{P}\big(x,t,\mathit{happens}(e,t)\big) \rightarrow \mathit{holds}\big(\mathit{vicinity}(x,e),t\big)\big]\Big)}{\mathbf{B}\big(a,t,\mathbf{B}\big(\gamma,t,\forall x,t,e : (x \neq \gamma) \rightarrow \big[\mathbf{P}\big(x,t,\mathit{happens}(e,t)\big) \rightarrow \mathit{holds}\big(\mathit{vicinity}(x,e),t\big)\big]\big)\big)}}{\mathbf{B}\ \Big(a,t,\mathbf{B}\ \Big(\gamma,t\Big[\neg \mathit{holds}\big(\mathit{vicinity}(b,e),t\big) \rightarrow \neg \mathbf{P}\big(b,t,\mathit{happens}(e,t)\big)\Big]\Big)\Big) \qquad \boxed{\phi_3}}}{\mathbf{B}\Big(a,t,\mathbf{B}\Big(\gamma,t,\neg \mathbf{P}(b,t,\mathit{happens}(e,t))\Big)\Big)}}{\mathbf{B}\big(a,t,\mathbf{B}(\gamma,t,\neg \mathbf{B}(b,t,\mathit{happens}(e,t)))\big) \equiv \boxed{\psi_2}}$$

From $\mathbf{C}_2^f$, we have:

$$D_{[\mathbf{C}\rightsquigarrow\mathbf{B}]}\ \dfrac{\mathbf{C}\ \big(t,\big[\mathbf{B}(\gamma,t,\mathit{happens}(e,t)) \land \mathbf{B}(\gamma,t,\neg \mathbf{B}\big(b,t,\mathit{happens}(e,t)\big)) \rightarrow \mathbf{S}(\gamma,b,t+\Delta,\mathit{happens}(e,t))\big]\big)}{\mathbf{B}(a,t,\big[\mathbf{B}(\gamma,t,\mathit{happens}(e,t)) \land \mathbf{B}(\gamma,t,\neg \mathbf{B}\big(b,t,\mathit{happens}(e,t)\big)) \rightarrow \mathbf{S}(\gamma,b,t+\Delta,\mathit{happens}(e,t))\big]\ ) \equiv \boxed{\psi}}$$

Using the above derived, $\boxed{\psi}$ and using $I_{\mathbf{B}}$:

$$I_{\mathbf{B}}\ \dfrac{\boxed{\psi} \qquad \mathbf{B}\big(a,t,\mathbf{B}(\gamma,t,\mathit{happens}(e,t))\big) \equiv \boxed{\psi_1} \qquad \mathbf{B}\big(a,t,\mathbf{B}(\gamma,t,\neg \mathbf{B}(b,t,\mathit{happens}(e,t)))\big) \equiv \boxed{\psi_2}}{\mathbf{B}(a,t,\mathbf{S}(\gamma,b,t+\Delta,\mathit{happens}(e,t)))}$$

■

## 6. Cognitive Polysolid Framework

We now introduce the *cognitive polysolid famework* (CPF), a class of problems that we use for experiments. From the framework, we can generate specific *cognitive polysolid world instantiations* in which we then declare the number of blocks/solids, their properties, and how these blocks/solids can be moved, as well as any agents and their possible beliefs or knowledge.

The CPF subsumes the familiar "blocks world," described for instance in Nilsson (1980), and long used for reasoning and planning in purely extensional ways. Briefly, CPF gives us a physical and cognitive domain unlike the purely physical blocks world domain. (The formal logic used in Nilsson (1980) is purely extensional, as it's simply first-order logic.) Since the physical complexities of blocks world problems have been well explored (Gupta & Nau, 1991), and since this microworld has been has been used for benchmarking (Slaney & Thiébaux, 2001), we emphasize cognitive extensions of it.

A cognitive polysolid world instantiation contains some finite number of blocks and a table large enough to hold all of them. Each block is `on` one other object; that object can be another block or the table. A block is said to be `clear` if there is no block that is on top of it. To move the blocks, an agent can either `stack` (placing a block on the table on top of another block) or `unstack` (taking a block that is on top of another block and placing it on the table). Before stacking the blocks, both need to be clear; when unstacking, the top block must be clear beforehand. After stacking the blocks, the bottom block is then not clear, and after unstacking, it is then clear. Translating this description to the $\mathcal{DCEC}$, we add two additional sorts and a constant, as well as some new functions; our augmentation is shown below:

$$\begin{array}{ll}
\textsf{Surface} \sqsubset \textsf{Object} & \mathit{clear} : \textsf{Block} \rightarrow \textsf{Fluent} \\
\textsf{Block} \sqsubset \textsf{Surface} & \mathit{goal} : \textsf{Formula} \times \textsf{Number} \rightarrow \textsf{Formula} \\
\mathit{table} : \textsf{Surface} & \mathit{stack} : \textsf{Block} \times \textsf{Block} \rightarrow \textsf{ActionType} \\
\mathit{on} : \textsf{Block} \times \textsf{Surface} \rightarrow \textsf{Fluent} & \mathit{unstack} : \textsf{Block} \times \textsf{Block} \rightarrow \textsf{ActionType}
\end{array}$$

## 7. A Cognitive Polysolid World Simulation

We start with a very elementary cognitive polysolid world (though this work scales fine to larger numbers). We have three identical blocks, named $A$, $B$, and $C$, which all start on the table. This is represented in the $\mathcal{DCEC}$ as:

$$\begin{array}{ll}
\mathit{holds}(\mathit{on}(A, \mathit{table}), 0) & \mathit{holds}(\mathit{clear}(A), 0) \\
\mathit{holds}(\mathit{on}(B, \mathit{table}), 0) & \mathit{holds}(\mathit{clear}(B), 0) \\
\mathit{holds}(\mathit{on}(C, \mathit{table}), 0) & \mathit{holds}(\mathit{clear}(C), 0)
\end{array}$$

There are only two human agents, $h_1$ and $h_2$, who have knowledge about how the cognitive polysolid world works. Using this instantiation, we give the room two tasks to demonstrate its theory of mind as required by the constraints specified above. For both tasks, we will use the following sequence of events to configure the world for the two tasks:

1. $h_1$ and $h_2$ enter the room
2. $h_1$ moves block $A$ onto block $B$

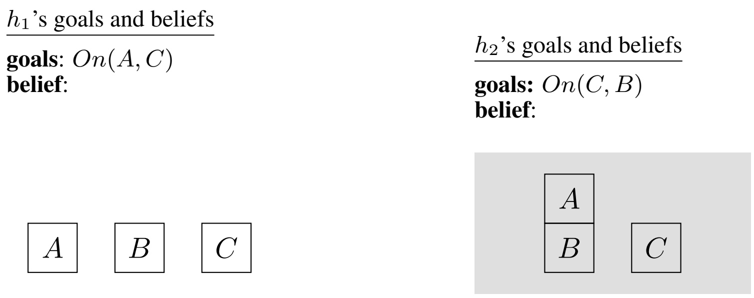

Figure 2: Visual representation of mental states of the agents ($h_2$'s state is grayed as he is not in the room).

3. $h_2$ adds the goal of block $C$ on block $B$
4. $h_2$ leaves the room
5. $h_1$ moves block $A$ to the table
6. $h_1$ removes the goal for block $C$ and adds the goal of block $A$ on block $C$
7. $h_1$ moves block $A$ onto block $C$
8. $h_2$ returns to the room
9. $h_2$ tries to move $A$ to the table.

For this simulation, all events and fluents inside the room are considered to be in the vicinity of agents within the room, and none of the events and fluents within the room are considered to be in the vicinity of agents outside the room when they happen or hold.

For the first portion of this task, we consider the world between steps 5 & 6. At this point, we wish to see where the room believes the blocks are, as well as where it believes that $h_1$ and $h_2$ think the blocks are, focusing primarily on block A. We ask the machine three questions, translating them into the $\mathcal{DCEC}$: *"Where does the CAIS/$h_1$/$h_2$ believe block A is?"*. We translate this question into three sentences in the $\mathcal{DCEC}$ which we can then pass down to ShadowProver to answer. For the first two questions, both the AI and the agent $h_1$ are in the room and can perceive where the block is, and thus have knowledge of its location. $h_2$ left the room at step 4 and missed the block being moved at step 5. Therefore, his knowledge of where the block is remains at what it was when he was in the room. We show the three statements below generated from ShadowProver that answers the above questions as well as show a visual representation of this answer in Figure 2:

$$\mathbf{B}(\mathit{cais}, t, \mathit{holds}(\mathit{on}(A, \mathit{table}), t))$$
$$\mathbf{B}(h_1, t, \mathit{holds}(\mathit{on}(A, \mathit{table}), t))$$
$$\mathbf{B}(h_2, t, \mathit{holds}(\mathit{on}(A, B), t))$$

For the second part of the task, we consider how the room responds after step 9 (say at $t_{10}$). The room believes that agent $h_2$ still believes that the goal is $On(C, B)$, and that it differs from the current goal of $On(A, C)$. The room then notifies $h_2$ showing a side-by-side comparison of the current goals of the room (and the plan to achieve it) and what his believed goals with respect to the room are (and the plan to achieve them). The system is able to do this, allowing for lag due to network calls, on average in around 748ms employing ShadowProver and Spectra multiple times.[14] Note that this scenario is an instantation of **Property 1** (Expectation of Usefulness) proved above. **Property 1** is instantiated with the event $e$ being setting a new goal: $e \equiv \mathit{setGoal}(On(A, C))$:

$$\begin{aligned}
&\mathbf{P}\big(h_1, t_6, \mathit{happens}\big(\mathit{setGoal}(On(A, C)), t_6\big)\big) \land \mathbf{P}\Big(h_1, t_6, \neg \mathit{holds}\big(\mathit{vicinity}(h_2, \mathit{setGoal}(On(A, C))), t_6\big)\Big) \\
&\qquad \rightarrow \mathbf{B}\Big(h_1, t_9, \mathbf{S}\big(\gamma, h_2, t_{10} \mathit{happens}(\mathit{setGoal}(On(A, C)), t_6)\big)\Big)
\end{aligned}$$

## 8. A Quick Summary

We give a quick summary before we delve into comparisons with prior/related work below.

(1) In Section 2.1 we explained why systems that are less expressive than quantified modal logic cannot model beliefs and other mental states with fidelity.

(2) We proposed in Section 3 a pair of informal requirements $\langle \mathcal{C}, \mathcal{I} \rangle$ that a CAIS should satisfy.

(3) In Section 4, we formalized these requirements in $\mathcal{DCEC}$, a quantified modal logic, resulting in:

$$\langle \mathbf{C}_1^f, \mathbf{C}_2^f, \mathbf{I}_1^f, \mathbf{I}_2^f, \mathbf{I}_3^f \rangle$$

(4) We then proved in Section 5 that these formal requirements lead to a desirable property: **Property 1** (a foundational property that can be used to build other such properties in future work).

(5) Finally, we briefly presented an implementation of a system that adheres to the formalization. We showed a scenario in a cognitive polysolid world where **Property 1** is useful and realized.

## 9. Prior/Related Work and Novelty

### 9.1 Microworlds

The environment for our case studies is, as we have said, *not* a blocks world. Blocks worlds are all deficient from the point of view of our research program, in significant part because the objects within them are devoid of propositional attitudes, and hence mere extensional logic suffices. A classic confirmatory presentation and lucid discussion of a classic blocks world can be found in Genesereth & Nilsson (1987); it will be seen in that discussion that the only objects in the microworld are inanimate and non-cognitive. The same will be seen in early, seminal attempts to formalize physical objects and processes, as for instance in Hayes (1978, 1985). In stark contrast, CPF is a member of a class of microworlds best lableled as *cognitive microworlds*; a *sine qua non* for a microworld being in this class is that some of the objects therein are cognitive agents that as

[14] The algorithmic details of the implementation are irrelevant and can be found in prior cited work. There is a deeper integration with the reasoner and planner that allows us to offload intensive reasoning tasks, necessitated by the principles given above, as planning tasks. Discussing this is beyond the scope of this paper. A video of this task can be viewed at http://mpeveler.com/cbwf-falsebelief.html

such have propositional attitudes and obligations, and attend, perceive, communicate at the level of human natural language, etc. Therefore a CPF includes instances in which the environment is populated by agents the representation of which requires *intensional* logic, and indeed quantified intensional logic.[15]

### 9.2 Unprecedented Expressivity

We have previously proved [e.g. in (Bringsjord & Govindarajulu, 2012)] that we cannot model even everyday propositional knowledge in non-intensional systems.[16] For the sort of problems we are interested in, minimally, first-order logic, married to multiple intensional operators, is required. For a simple instance of this point, consider that we need to have uncompromising representations of statements such as:

1. *"There is no one in the room who believes that no one is in the room."*
2. *"The organizer believed that the number of people in the room (7) was more than what was allowed, and hence had to ask some of the participants to leave."*

It is impossible to model these statements in systems that are not at once non-quantificational or non-intensional in nature. Even prior art such as the event calculus and the systems in Wooldridge (2002), which are sensitive to expressivity demands, are based on systems that are markedly less expressive than the quantified multi-operator modal calculus we use in the present paper.[17] Along the same line, from the formal point of view, the excellent, aforementioned Pearce et al. (2014) allows nested belief operators, but their scope does not include unrestricted quantification, rather sub-formulae in zero-order logic; and the other intensional operators, epistemic and otherwise, are absent. The false-belief task was provably solved by the logicist AI system specified and implemente3d in Arkoudas & Bringsjord (2009a), and even infinitary cognitive challenges are met in Arkoudas & Bringsjord (2004), but this work is essentially a proper subset and ancestor of the richer formalism and technology brought to bear in our attack on CAISs.

### 9.3 Rejection of Fixed Logics

Our work is not based on a particular, standard logic, such as are brought to bear in the epistemic case in the likes of (Moore, 1985) and (Fagin et al., 2004), or in work that directly appropriates a

[15] In further contrast, CPF can include irreducibly visual entities the representation of which requires heterogeneous logic Barwise & Etchemendy (1995); Arkoudas & Bringsjord (2009b), which none of the prior/related researchers in the logicist tradition referred to in the present section ever investigated, since they worked/work on straight symbolic systems, with no homomorphic representations to be found. Moreover, even in the case of non-mental objects, arbitrary polysolids are allowed, and these solids can be in motion through time. The case studies discussed in the present paper don't employ the full available range of entities allowable in a CPF; and formal specification of CPF and the broader category of cognitive microworlds is out of scope here.

[16] The founders of modern AI, many of whom were logicits, all came to the field from mathematical logic, which is by definition extensional, not intensional; see e.g. (Ebbinghaus et al., 1994) for discussion. The irony here is that Leibniz can be viewed as the primogenitor of logicist AI, and he invented *both* modern extensional and intensional formal logic.

[17] For a more detailed discussion, please see the appendix in Govindarajulu & Bringsjord (2017a).

logic-programming base and syntax, since such bases are invariably built atop extensional resolution. Indeed, it's in part precisely because of the inadequacy of "off the shelf" logics [such as those in the well-known $\Diamond$-$\Box$ ontology for modal/epistemic logics] that motivated the creation of cognitive calculi in the first place. As seen above, cognitive calculi are a space of formal systems that are composed of a typed signature $\Sigma$ (which in turn include an alphabet, formal grammar, and type information), along with a tailor-made, easily adjustable collection of inference schemata. Hence, given cognitive calculus is unlikely to correspond to any fixed, rigid logic; moreoever, any use of a informal syntax renders the work based upon this use radically different than our specification of signatures. Almost all prior work in AI in the logicist tradition is based on off-the-shelf logics, with standard, long-standing inference rules. In fact, the early, seminal work done in this tradition by Newell, Simon, McCarthy, and Hayes was based on exploitation of the propositional and predicate calculi. Our approach is dramatically different, in that *any* collection of natural (where 'natural' is a here a technical term, one used in e.g. 'natural deduction) inference schemata can be created and used, and immediately implemented via corresponding adjustments in ShadowProver. Note as well that inference schemata can be, and often in our work are, meta-logical.[18] This enables what appears to be unprecedented flexibility.

### 9.4 Rejection of Standard Semantics/Models in Favor of Proof-Theoretic Semantics

Here is a telling quote from early work by Hayes:

> The ability to interpret our axioms in a possible world, see what they say and whether it is true or not, is so useful that I cannot imagine proceeding without it. . . . The main attraction of formal logics as representational languages is that they have very precise model theories, and the main attraction of first-order logic is that its model theory is so simple, so widely applicable, and yet so powerful. (Hayes 1985, 10)

This quote is telling because our cognitive calculi are purely inference-theoretic, and are traceable in this regard directly back to the advent of proof-theoretic semantics Prawitz (1972), which eschews semantics of the sort that Hayes venerates. A cognitive calculus has no provided model theory for its extensional levels, and rejects any use for instance of possible worlds to provide non-inferential semantics for modal operators, a rejection that can be traced back to an early proof by Bringsjord (1985) that standard set-theoretic unpackings of possible worlds lead to absurdity.

### 9.5 What About Bayesian Approaches?

We have of course placed the underlying technical content of our AI work on CAISs within the logicist tradition.[19] What about Bayesian approaches to reaching the goals we have for smart rooms and in particular CAISs? How does such work relate technically to cognitive calculi and the reasoning

[18] E.g., a schema might say that if $\phi$ is provable in less than $k$ steps, then an agent should believe $\phi$.

[19] We have for economy herein gone "light" on the historical trajectory of this approach (which is easily traceable to Leibniz, and even to Aristotle) to both modeling and mechanizing human-level cognition. For an overview of *all* of AI, including from an historical point of view, and covering Bayesianism in the field as well, see Bringsjord & Govindarajulu (2017). Dedicated coverage of the logicist approach in AI and computational cognitive science can be found here: (Bringsjord, 2008b,a).

and planning systems that work symbiotically with them (all covered, of course, above)? We do not have space for a sustained answer to these questions, and therefore opt for stark brevity via but two points: One, since Bayesian work in AI is of necessity based on underlying formal languages of the extensional and simple sort (e.g., propositional and predicate calculi with the key function symbol *prob* whose range is $[0,1]$) used to specify at least parts of the probability calculus going back to Kolmogorov, and since these formal languages are painfully inexpressive relative to cognitive calculi, Bayesian formalisms and technology built atop them are inadequate for reaching the goals we have set.[20] Two: Bringsjord and Govindarajulu subsume probability into a proof-theoretic approach based on multi-valued cognitive calculi, where the values are likelihood values, or — as they are sometimes known — strength factors. For details, see (Govindarajulu & Bringsjord, 2017b).

### 9.6 Novelty?

While intelligent and immersive rooms have been designed and built by researchers for decades, there has not been any formal, rigorous work characterizing an intelligent room and how it differs from other AI-infused environments. We offer the first such characterization of an intelligent room that has cognitive abilities: a CAIS powered by AI operating on the strength of cognitive calculi, which are in turn empowered by implemented reasoning and planning technology, where specific requirements must be met. We have also shown that a room that satisfies these requirements can have a useful property that we term *expectation of usefulness*. One advantage of our formalization is that we have used a quantified multi-operator modal logic that has been previously used to model higher-level cognitive principles. This opens up possibilities for principled integration of other cognitive abilities in the future.

## 10. Conclusion

We now quickly summarize our contributions and present promising future lines of work. Our primary contribution is in the creation of an overseeing AI that is capable of tracking participants' mental states as they operate within a CAIS. This AI is capable of using this information to reason and plan assisting the participants in completing a task and tracking information as well as generate explanations for its actions. As part of this system, we give a definition for a concrete framework that can be used to generate future tests of increasing complexity for the system. We give the necessary machinery via syntax and sorts to use this framework as well as show a base implementation that was derived from the framework. Within this implementation, we show two tasks that can be solved via the the overarching AI that encompasses a CAIS and has a sufficient definition of theory of mind.

Future work will be on further empowering the overarching AI, in creating additional domains of work, and to create an overarching formal definition for a *cognitive and immersive framework*. To extend the AI framework, we are considering adding partial satisfaction planning capabilities (Van Den Briel et al., 2004; Smith, 2004) to the Spectra planner so that it can reason about goals that

[20] Even inductive logicians of the first rank readily acknowledge that the probability of such propositions as that Jones believes that Smith believes that Jones believes that there are exactly two properties shared between Black and Smith are inexpressible in Bayesianism; e.g. see (Fitelson, 2010).

conflict with each other in a utility-centric framework. In addition, we currently allow agents to add goal priority explicitly. However, in a real group discussion, priority of a given goal may shift implicitly or by recognizing the particular belief states of the agents about a given goal. To improve upon this, we aim to incorporate research of social choice theory and ranking data via such methods as discussed in Xia (2017).

The *cognitive and immersive framework* would form the basis for any of our cognitive frameworks, which includes the CPF. From the core cognitive definition of the CPF, we hope to expand our system to other domains that would benefit from the cognitive capabilities demonstrated here. For one such domain, the presented work can be very useful in business negotiation scenarios like contracts management or using software where multiple parties are involved and can have different vantage points (mental models) for discussion. There has been some work on analyzing contracts for identifying gaps (Desai et al., 2008) and terms of conditions for software services (Vukovic et al., 2014) but they do not consider parties' mental models. Finally, we aim to further enhance and refine the CPF such that it can be used to capture and reason about more complex domains than the presented blocks world, such as described in Barker-Plummer et al. (2017); Johnson et al. (2016) and how agents may interact with them.

## 11. Acknowledgments

We are grateful for both IBM and Rensselaer support of the research reported on herein, via both the CISL and AIRC programs. Our gratitude extends to both ONR and AFOSR, support from which has among other things in part enabled the invention, implementation, and refinement of ShadowProver and Spectra. Four anonymous referees provided helpful comments, and we deeply appreciate their insights. Trenchant, precise analysis provided by Pat Langley proved to be invaluable, and we are indebted.

## References

Arkoudas, K., & Bringsjord, S. (2004). Metareasoning for multi-agent epistemic logics. *Proceedings of the Fifth International Conference on Computational Logic In Multi-Agent Systems (CLIMA 2004)* (pp. 50–65). Lisbon, Portugal.

Arkoudas, K., & Bringsjord, S. (2008). Toward formalizing common-sense psychology: An analysis of the false-belief task. *Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence* (pp. 17–29). Springer-Verlag.

Arkoudas, K., & Bringsjord, S. (2009a). Propositional attitudes and causation. *International Journal of Software and Informatics*, *3*, 47–65.

Arkoudas, K., & Bringsjord, S. (2009b). Vivid: An AI framework for heterogeneous problem solving. *Artificial Intelligence*, *173*, 1367–1405.

Barker-Plummer, D., Barwise, J., & Etchemendy, J. (2017). *Logical reasoning with diagrams and sentences: Using hyperproof*. CSLI lecture notes. CSLI Publications/Center for the Study of

Language & Information.

Barwise, J., & Etchemendy, J. (1995). Heterogeneous logic. In J. Glasgow, N. Narayanan, & B. Chandrasekaran (Eds.), *Diagrammatic reasoning: Cognitive and computational perspectives*, 211–234. Cambridge, MA: MIT Press.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*, 227–247.

Bringsjord, S. (1985). Are there set-theoretic worlds? *Analysis*, *45*, 64.

Bringsjord, S. (2007). Offer: One billion dollars for a conscious robot. If you're honest, you must decline. *Journal of Consciousness Studies*, *14*, 28–43.

Bringsjord, S. (2008a). Declarative/logic-based cognitive modeling. In R. Sun (Ed.), *The Handbook of Computational Psychology*, 127–169. Cambridge, UK: Cambridge University Press.

Bringsjord, S. (2008b). The logicist manifesto: At long last let logic-based AI become a field unto itself. *Journal of Applied Logic*, *6*, 502–525.

Bringsjord, S. (2010). Meeting Floridi's challenge to artificial intelligence from the knowledge-game test for self-consciousness. *Metaphilosophy*, *41*, 292–312.

Bringsjord, S., Bello, P., & Govindarajulu, N. (2018). Toward axiomatizing consciousness. In D. Jacquette (Ed.), *The bloomsbury companion to the philosophy of consciousness*, 289–324. London, UK: Bloomsbury Academic.

Bringsjord, S., & Govindarajulu, N. S. (2012). Given the web, what is intelligence, really? *Metaphilosophy*, *43*, 361–532.

Bringsjord, S., & Govindarajulu, N. S. (2017). Artificial Intelligence. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. From `https://plato.stanford.edu/entries/artificial-intelligence`.

Bringsjord, S., Licato, J., Govindarajulu, N., Ghosh, R., & Sen, A. (2015). Real robots that pass tests of self-consciousness. *Proccedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 498–504). New York, NY: IEEE.

Brooks, R. A. (1997). The intelligent room project. *Proceedings of Second Internal Conference on Cognitive Technology, 1997. Humanizing the Information Age.* (pp. 271–278). IEEE, Aizu-Wakamatsu City, Japan: IEEE Computer Society Press.

Chakraborti, T., Briggs, G., Talamadupula, K., Zhang, Y., Scheutz, M., Smith, D., & Kambhampati, S. (2015). Planning for serendipity. *Proceedings of 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems2015 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 5300–5306). IEEE, Hamburg, Germany: IEEE.

Chakraborti, T., Talamadupula, K., Dholakia, M., Srivastava, B., Kephart, J. O., & Bellamy, R. K. E. (2017). Mr. Jones – Towards a proactive smart room orchestrator. *AAAI Fall Symposium on Human-Agent Groups*.

Chan, M., Estève, D., Escriba, C., & Campo, E. (2008). A review of smart homes—present state and future challenges. *Computer methods and programs in biomedicine*, *91*, 55–81.

Cirillo, M., Karlsson, L., & Saffiotti, A. (2009). Human-aware task planning for mobile robots. *Proceedings of 2009 International Conference on Advanced Robotics* (pp. 1–7). IEEE.

Coen, M. H., et al. (1998). Design principles for intelligent environments. *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence* (pp. 547–554). Madison, Wisconsin, USA: AAAI Press.

Desai, N., Narendra, N. C., & Singh, M. P. (2008). Checking correctness of business contracts via commitments. *Proceedings of 7th Int. Conf. on Autonomous Agents and Multiagent Systems*.

Divekar, R. R., Peveler, M., Rouhani, R., Zhao, R., Kephart, J. O., Allen, D., Wang, K., Ji, Q., & Su, H. (2018). Cira—an architecture for building configurable immersive smart-rooms. *To appear in Proceedings of Intellisys 2018*.

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

Fagin, R., Halpern, J., Moses, Y., & Vardi, M. (2004). *Reasoning about knowledge*. Cambridge, MA: MIT Press.

Fitelson, B. (2010). Pollock on Probability in Epistemology. *Philosophical Studies*, *148*, 455–465.

Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology*, *15*, R644–R645.

Genesereth, M., & Nilsson, N. (1987). *Logical foundations of artificial intelligence*. Los Altos, CA: Morgan Kaufmann.

Gentzen, G. (1964). Investigations into logical deduction. *American Philosophical Quarterly*, *1*, 288–306.

Govindarajulu, N., & Bringsjord, S. (2017a). On automating the doctrine of double effect. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 4722–4730). Melbourne, Australia: International Joint Conferences on Artificial Intelligence.

Govindarajulu, N. S., & Bringsjord, S. (2017b). Strength factors: An uncertainty system for quantified modal logic. *Proceedings of the IJCAI Workshop on "Logical Foundations for Uncertainty and Machine Learning* (pp. 34–40). Melbourne, Australia.

Gupta, N., & Nau, D. S. (1991). Complexity results for blocks-world planning. *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2* (pp. 629–633). Anaheim, California: AAAI Press.

Hayes, P. (1978). The Naïve physics manifesto. In D. Mitchie (Ed.), *Expert systems in the microeletronics age*, 242–270. Edinburgh, Scotland: Edinburgh University Press.

Hayes, P. J. (1985). The Second naïve physics manifesto. In J. R. Hobbs & B. Moore (Eds.), *Formal theories of the commonsense world*, 1–36. Norwood, NJ: Ablex.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. B. (2016). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, *abs/1612.06890*. From `http://arxiv.org/abs/1612.06890`.

Kim, J., & Shah, J. (2017). Towards intelligent decision support in human team planning. *AAAI Fall Symposium on Human-Agent Groups*.

Langley, P., Barley, M., Choi, D., Katz, E., & Meadows, B. (2016). Goals, utilities, and mental simulation in continuous planning. *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems*.

Moore, R. (1985). Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, *25*, 75–94.

Mueller, E. (2014). *Commonsense reasoning: An event calculus based approach*. Morgan Kaufmann, second edition.

Nilsson, N. J. (1980). *Principles of artificial intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pearce, C., Meadows, B., Langley, P., & Barley, M. (2014). Social planning: Achieving goals by altering others' mental states. *Proceedings of the 28th AAAI Conference on Artificial Intelligence* (pp. 402–408). Palo Alto, CA: AAAI.

Prawitz, D. (1972). The philosophical position of proof theory. In R. E. Olson & A. M. Paul (Eds.), *Contemporary philosophy in scandinavia*, 123–134. Baltimore, MD: Johns Hopkins Press.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *4*, 515–526.

Sengupta, S., Chakraborti, T., Sreedharan, S., & Kambhampati, S. (2017). RADAR - A proactive decision support system for human-in-the-loop planning. *AAAI Fall Symposium on Human-Agent Groups*.

Slaney, J., & Thiébaux, S. (2001). Blocks world revisited. *Artificial Intelligence*, *125*, 119 – 153.

Smith, D. E. (2004). Choosing objectives in over-subscription planning. *Proceedings of the Fourteenth International Conference on International Conference on Automated Planning and Scheduling* (p. 393).

Talamadupula, K., Briggs, G., Chakraborti, T., Scheutz, M., & Kambhampati, S. (2014). 2014 ieee/rsj international conference on intelligent robots and systems. *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2957–2962). Chicago, IL: IEEE.

Van Den Briel, M., Sanchez, R., Do, M. B., & Kambhampati, S. (2004). Effective approaches for partial satisfaction (over-subscription) planning. *Proceedings of the Nineteenth National Conference on Artificial Intelligence* (pp. 562–569). San Jose, CA: AAAI Press.

Vukovic, M., Laredo, J., & Rajagopal, S. (2014). API terms and conditions as a service. *Proceedings of 2014 Services Computing Conference*.

Wooldridge, M. (2002). *An introduction to multi agent systems*. Cambridge MA: MIT Press.

Xia, L. (2017). Improving group decision-making by artificial intelligence. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 5156–5160). Melbourne, Australia.