Toward Formalizing Teleportation of Pedagogical Artificial Agents

John Angel  
*Rensselaer Polytechnic Institute*, Angelson1992@gmail.com

Naveen S. Govindarajulu  
*Rensselaer Polytechnic Institute*, naveensundarg@gmail.com

Selmer Bringsjord  
*Rensselaer Polytechnic Institute*, selmerbringsjord@gmail.com

Follow this and additional works at: https://digitalcommons.montclair.edu/eldj

Part of the Curriculum and Instruction Commons, and the Online and Distance Education Commons

**Recommended Citation**

Available at: https://digitalcommons.montclair.edu/eldj/vol6/iss1/2

This Article is brought to you for free and open access by Montclair State University Digital Commons. It has been accepted for inclusion in The Emerging Learning Design Journal by an authorized editor of Montclair State University Digital Commons. For more information, please contact digitalcommons@montclair.edu.
Toward Formalizing Teleportation of Pedagogical Artificial Agents

Cover Page Footnote
All authors of this paper are members of the Rensselaer AI & Reasoning (RAIR) Lab. Furthermore, the RAIR lab is affiliated with both the Department of Computer Science & Department of Cognitive Science at Rensselaer Polytechnic Institute

This article is available in The Emerging Learning Design Journal: https://digitalcommons.montclair.edu/eldj/vol6/iss1/2
Toward Formalizing Teleportation of Pedagogical Artificial Agents
John Angel*, Naveen S. Govindarajulu, Selmer Bringsjord
Rensselaer Polytechnic Institute
*angelj3@rpi.edu
December 5, 2018

ABSTRACT
Our paradigm for the use of artificial agents to teach requires among other things that they persist through time in their interaction with human students, in such a way that they “teleport” or “migrate” from an embodiment at one time t to a different embodiment at later time t’. In this short paper, we report on initial steps toward the formalization of such teleportation, in order to enable an overseeing AI system to establish, mechanically, and verifiably, that the human students in question will likely believe that the very same artificial agent has persisted across such times despite the different embodiments. The system achieves this by demonstrating to the students that different embodiments share one or more privileged beliefs that only one single agent can possess.

Keywords: Adaptive/Personalized Learning, Artificial Intelligence, Mobile Learning

INTRODUCTION
Our paradigm for the use of artificial agents to teach requires among other things that they persist through time in their interaction with human students, in such a way that they “teleport” or “migrate” from an embodiment at one time t, to a different embodiment at a later time. In this article, we report on initial steps toward the formalization of such teleportation, in order to enable an overseeing AI system to establish, mechanically, and verifiably, that the human students will likely believe that the very same artificial agent has persisted across such times despite the different embodiments.

The plan for the paper is straightforward, and as follows. After encapsulating our paradigm for the deployment of artificial agents in service of learning, and taking note of the fact that the “teleportation”/“migration” problem has hitherto been treated only informally, we convey the kernel of our approach to formalizing agent teleportation between different embodiments, then formalize this kernel in order to produce an initial simulation, and wrap up with some final remarks.

Our Paradigm & Teleportation
A crucial part of our novel paradigm for artificial agents that teach is the engineering of a class of AIs, crucially powered by cognitive logics, able to persist through days and weeks in their interaction with the humans whose education is to be thereby enhanced. The artificial agents in our paradigm are able to seamlessly “teleport” between heterogeneous environments in which a human learner may find herself as time unfolds; this capacity is intended to provide a continuous educational experience to the human student, and offers the possibility of human-machine friendship.

In short, our agents need to be “teleportative.” This means that the agent should be usable in multiple hardware environments by a user, such that the user has the impression of a continuous, uninterrupted interaction with the very same agent. This helps to reinforce the possibility of a persistent, trusting relationship between human and machine. See Figure 1 below for one implemented incarnation of such a system: TIPPAE (Teleportative Intelligent Persistent Personalized Agents for Education). As can be seen in the figure, other than possibly sharing names, there is no explicit information that indicates that the same agent persists across the interactions. Our contribution in this article is aimed at addressing this issue.
Definitions: Some definitions are in order before we go any further. An agent (or person) is either a human or any artificial system of sufficient intelligence. Our usage is not different from standard uses of the word “agent” in the AI literature (Russell & Norvig 2009). An embodiment (or manifestation, or presentation) is any physical or virtual interface for an agent. A single artificial agent can have multiple embodiments.

We highlight below the challenge we seek to solve.

**Challenge:** Given that the same agent \(a\) can have different physical embodiments \((m_1, \ldots, m_n)\), in different physical and virtual educational environments, how do we convince a student \(u\) interacting with agent \(a\), that despite differences in embodiments, the student \(u\) is dealing with the same agent \(a\)?

Briefly, our solution leverages cognition and is summarized below.

**Solution Summary:** Since embodiments vary, the agent \(a\) has to convince \(u\) that it is the same agent based on demonstrating to \(u\) one or more personal beliefs about \(u\) that all the embodiments, and only the embodiments of \(a\) possess.

**PRIOR ACCOUNTS OF TELEPORTATION OF ARTIFICIAL AGENTS**

There is some excellent and interesting prior work on teleporting artificial agents. Some explore how the consistency of a migrating agent’s memory affects a user’s perception of a continuous identity (Aylett et al., 2013) and have suggested that migrating the long-term memories of an agent could have a stronger effect than migrating short-term memories, something that our paradigm is uniquely positioned to explore. Others shed light on visual cues useful for convincing users of an agent’s teleportation (Koay, Syrdal, Walters, & Dautenhahn, 2009) by illustrating how cues imply both a connection between embodiments and the migration of the agent; a simple example of this could be a bar on the previous embodiment slowly emptying while a bar on the next embodiment fills in, to enhance the impression of teleportation. In addition, progress has been made toward the design of migrating agents (Hassani & Lee, 2014) and testing real-world implementations of such agents (Gomes et al., 2011). All of these works help to explore, flesh out, and define what a teleportative agent should be; unfortunately for our purposes, the prior art is informal. Our goal is to capture teleportation formally.
and on the strength of that formalization to enable an overseeing AI system to prove, or minimally justify rigorously, that the teleportation in question is indeed believable.

**THE KERNEL OF THE FORMALIZATION**

In the longstanding quasi-technical literature on personal identity in philosophy, there is a strong tradition of trying to work out a rigorous account of when person (or agent) $p_1$ at $t_1 (= p_{t1})$ is identical with person (or agent) $p_2$ on the basis of shared memories between $p_{t1}$ and $p_{t2}$. More concretely, for our problem, as mentioned, a given agent or person can have multiple embodiments across time. The goal here is to determine, in some rigorous manner, whether two different embodiments are of the same agent.

Simple schemes such as the embodiment of the agent sharing the same name or appearance might not always work. Sharing of names is unreliable since there might be more than one virtual agent with the same name. For example, is Apple’s Siri on two different devices the same agent? The same worry infects any such proposal as that sharing appearance will settle matters. Furthermore, sharing appearances might not always even be possible. For instance, in the example in Figure 1, the embodiment in the first instance has no physical representation and in the second instance, a robot represents the agent.

The above argument demonstrates that we need to have a deeper model of a virtual agent being the same or different across different embodiments. The gist of our scheme, reflective of the line of thinking on personal identity in philosophy mentioned above, is that the embodiments can be considered to be the same if they share certain privileged beliefs. These beliefs are ones that only a single agent could possibly access.

The goal of our initial formalization here is to build a system that can find a proof for when it believes that a student believes two embodied agents are the same. The system can conclude that the student believes two embodiments to be the same if the system can find a proof that it believes that the student believes that the two embodiments have a privileged belief $\beta$ at specific times that cannot be believed by more than one agent. If the system fails to find such a proof or argument, then the system can take corrective actions to make it more explicit to the human that the embodiments are the same. Note that formalization requires the system to understand beliefs of agents which might themselves be about beliefs of other agents (and so on).

**INITIAL FORMALIZATION & SIMULATION**

The requirement that the system understand the student’s beliefs about other embodied agents’ beliefs implies that we need to have a sufficiently expressive system. BDI logics (belief/desire/intentions) have a long tradition of being used to model such agents (Wooldridge, 2002) with deep beliefs.

**Formal system**

For our formalization, we use a system that is in the general tradition of such logics. We specifically use the formal system *DCEC* (deontic cognitive event calculus) used in (Govindarajulu & Bringsjord 2017). DCEC has a well-defined syntax and inference system; see Appendix A of (Govindarajulu & Bringsjord, 2017a) for a full description of the technical details of the system. The inference system in DCEC is based on natural deduction (Gentzen, 1935), which is an inference system commonly used by practicing mathematicians and by educators in logic.

This calculus itself is a first-order modal logic (Boo1os, et al., 2002) and belongs to a family of cognitive calculi. Cognitive calculi are formal systems designed to model and automate multiple agents with beliefs, desires, intentions, and other cognitive states, interacting over time. Cognitive calculi include the event calculus (Mueller, 2014), a system for reasoning over the physical world and commonsense phenomena. More specifically, DCEC is designed to model ethical principles. For instance, DCEC has been used previously
by Govindarajulu and Bringsjord (2017a) to formalize and automate versions of the *doctrine of double effect*, an ethical principle with deontological and consequentialist components. Cognitive calculi have also been used to formalize and automate highly cognitive reasoning processes, such as the false-belief task (Arkoudas and Bringsjord, 2008) and *akrasia* (succumbing to temptation to violate moral principles) (Bringsjord et al., 2014). Arkoudas and Bringsjord (2008) introduced the general family of cognitive event calculi to which DCEC belongs, through their formalization of the false-belief task. While describing the calculus is beyond the scope of this article, we give an example representation of a complex belief represented in the calculus in Table 1. The logical operator “\(B\)” below represents a belief.

Although it is possible to install YOURLS as a subdirectory on a website, it is counterintuitive if the goal is to shorten URLs. The reason being, is that an additional subdirectory will create a longer URL. For example, if a site’s domain is “www.domain.edu/YOURLS,” the link will be much more to input into a device.

**Table 1.** Example representation of information in DCEC (deontic cognitive event calculus).

<table>
<thead>
<tr>
<th>Language</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>English statement</td>
<td><em>John believes now that Mary believes that it is raining now.</em></td>
</tr>
<tr>
<td>DCEC Representation</td>
<td>(B(\text{john}, \text{now}, B(\text{mary}, \text{now}, holds(\text{raining}, \text{now})))))</td>
</tr>
</tbody>
</table>

**Simulation**

The simulation is set up as a reasoning problem from a set of given assumptions to a goal (see Figure 4). In the formalization shown below in Figure 4, the system believes that the student believes two embodiments to have the same identity if the embodiments at different times believe *some personal object of the student to have the same property* (Assumption A4 in Figure 4). For instance, assume that the student’s watch is a personal object. At time \(t_1\), we have embodiment \(a\) believing that the watch is stopped, and at time \(t_2\) we also have embodiment \(b\) believing the same. From these assumptions, the system can derive that the student believes that the embodiments are the same. See Figure 3 for an example.
different times, then the two embodiments are of the same agent.

5. **Goal:** Finally, the human reasons that embodiment a and embodiment b are of the same agent.

Note that the above reasoning takes place in the mind of the TIPPAE agent. If the agent can successfully prove the above goal from a set of assumptions that it has access to, then it can conclude that the human believes that its two different embodiments are of the same agent.

**TIPPAE Revisited with Shared Beliefs across Embodiments**

How might the example scenario we showed in Figure 1 with TIPPAE be changed to accommodate the formal model presented above? In at least two interactions, TIPPAE needs to convey to the student that it believes one or more things about the student that no other agent can believe. Trivially, it can be the student’s state of progress in the domain being taught. The agent can also remember particular issues that the student might be facing in the learning task. For some teaching problems, such as the math problem shown in Figure 1, it might be easier to identify such beliefs and attributes than in other teaching problems. Unrelated to the learning task, as shown in the simulation above in Figure 3, the belief can be about an event or object not related to the learning task at hand. Somewhat relatedly, in the domain of user profiling (for example as in Middleton et al., 2004), statistical information about users is gathered en masse. User profiling systems, in general, do not make use of individual pieces of information about users (though such information might be gathered). While user profiling systems could help in making TIPPAE even more personalized to start with, TIPPAE would need to gather specific beliefs about an individual student and demonstrate to the student that TIPPAE has those beliefs.

**ETHICAL AND PRIVACY ISSUES**

One obvious issue is the privacy of the student when a teleportative agent seeks to learn some information that is unrelated to the learning task (as shown in the simulation above). This can be handled by regimenting the agent, by forbidding it from acquiring any information about the student that is not public. A rough sketch of such a condition cast in the language of DCEC is shown below in Table 2. The “F” operator in the example below represents that it is forbidden to do something. The “B” operator stands for belief as before.

<table>
<thead>
<tr>
<th>Language</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>English statement</td>
<td>TIPPAE is forbidden to believe any nonpublic information about the student</td>
</tr>
<tr>
<td>DCEC Representation</td>
<td>F(B(tippae, info)) ∧ belongs(info, student) ∧ ¬B(public, info)</td>
</tr>
</tbody>
</table>

**CONCLUDING REMARKS & NEXT STEPS**

We readily admit to having only taken initial steps toward the formalization of teleportation for artificial agents. The simulation we have presented does seem to indicate to us that things are scalable — but of course only time and experimentation will tell. Finally, it’s important to note that we haven’t herein sought to address the educational efficacy of our approach, nor the specific learning value of persistent teaching agents across embodiments.

Several possible venues of research exist in this direction. **Vertical studies** will focus on a student progressing through a sequence of increasingly harder topics in a class or subject (e.g. a sequence of topics \( T_1, T_2, \ldots, \) in trigonometry). **Horizontal studies** will focus on a student learning and applying a topic in one or more subjects (a student learning a topic \( T \) in trigonometry and applying \( T \) in a physics class). Finally, TIPPAE and its use in a group context, such as helping different members of a study group based on how advanced they are, and taking into account their interactions with others, is another rich area of research that can be explored. For instance, if a TIPPAE agent knows that a student \( s_1 \) has difficulty with topic \( T \) but another student \( s_2 \) has mastered it, the agent can suggest \( s_1 \) seek help on \( T \) from \( s_2 \).

While ethical and privacy concerns exist, the strength of the underlying formal system in modeling complex principles can possibly help us address these concerns. Particularly, DCEC has been used to model the **doctrine of double effect**, a principle that is used by (both formally ethically trained and untrained) humans to handle a number of longstanding moral dilemmas (Govindaraju & Bringsjord 2017a). DCEC has also been used to model other ethical theories and principles.
We believe that any significant ethical or privacy concerns that might have to be handled by the TIPPAE agent itself, can be handled by the ethical principles that have already been modeled in DCEC.

REFERENCES


This work is licensed under a Creative Commons Attribution-Non-Commercial-No Derivatives 4.0 International License [CC BY-NC-ND 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/)

This article is being published as a proceeding of the 2018 Emerging Learning Design Conference (ELDc 2018).