Computational Thought Experiments for a More Rigorous Philosophy and Science of the Mind

Iris Oved (irisoved@gmail.com)
Independent Scholar, The Paradox Lab
San Francisco, CA 94115 USA

Nikhil Krishnaswamy (nkrishna@colostate.edu)
Department of Computer Science, Colorado State University
Fort Collins, CO 80523 USA

James Pustejovsky (jamesp@brandeis.edu)
Department of Computer Science, Brandeis University
Waltham, MA 02453 USA

Joshua Hartshorne (joshua.hartshorne@hey.com)
Department of Psychology, Boston College
Chestnut Hill, MA 02467 USA

Abstract

We offer philosophical motivations for a method we call *Virtual World Cognitive Science* (VW CogSci), in which researchers use virtual embodied agents that are embedded in virtual worlds to explore questions in the field of Cognitive Science. We focus on questions about mental and linguistic representation and the ways that such computational modeling can add rigor to philosophical thought experiments, as well as the terminology used in the scientific study of such representations. We find that this method forces researchers to take a god's-eye view when describing dynamical relationships between entities in minds and entities in an environment in a way that eliminates the need for problematic talk of belief and concept *types*, such as *the belief that cats are silly*, and *the concept CAT*, while preserving belief and concept *tokens* in individual cognizers' minds. We conclude with some further key advantages of VW CogSci for the scientific study of mental and linguistic representation and for Cognitive Science more broadly.

Keywords: philosophy, methods; concepts; virtual worlds; embodiment; grounding; mental representation; AGI, LLMs.

Introduction

This paper offers philosophical motivations for a method we call *Virtual World Cognitive Science* (VW CogSci), in which researchers use virtual embodied agents that are embedded in virtual worlds to explore questions in the field of Cognitive Science. We offer a general defense of this method and then focus on the study of mental and linguistic representations.

First we recall some of the virtues of computational modeling in general, and then consider reasons to treat cognition as *embodied* (to include sensory receptors and motion actuators) and *embedded* (in a world to be sensed, represented, and acted upon). We then show how the method of VW CogSci allows researchers to go beyond the study of *actual* minds and how they operate in *actual* environments, to the study of various *possible* minds and how they might perform in various *possible* environments, with few practical and ethical constraints.

We then turn specifically to the study of mental and linguistic representation for the remainder of the paper. We summarize some of the persistent philosophical puzzles about beliefs and concepts, which many theorists either ignore or go through painful contortions to accommodate. These include puzzles from Saul Kripke (1979) and Hilary Putnam (1975). We then argue that similar puzzles arise not only in far-out philosophical scenarios, but for the common, everyday development of concepts and beliefs, and thus must be taken seriously in a study of cognition. We illustrate this by describing 'computational thought experiments' in which two young children hear and then see what in fact are some coyotes and wolves, and try to carve up the categories in their environment to best explain their experiences. We show how VW CogSci dissolves the philosophical puzzles by providing a god's-eye view that eliminates the need for problematic talk of belief and concept *types*, such as *the belief that cats are silly*, and *the concept CAT*, while preserving belief and concept *tokens* in individual cognizers' minds. We also show how the method allows for a more rigorous science of the mind via the simulation of complex dynamical relationships between mental entities and entities in an environment.

Related Work

Technologies in recent decades have enabled a method we call *Virtual World Cognitive Science* (VW CogSci), in which researchers build virtual agents in virtual worlds to test by simulation hypotheses about how minds and environments might interact. This is a method that has been used in cognitive robotics (Brockman et al., 2016; Oudeyer, Kaplan, & Hafner, 2007; Lungarella, Metta, Pfeifer, & Sandini, 2003; Asada et al., 2009; Cangelosi & Schlesinger, 2015; Mattern, López, Ernst, Aubret, & Triesch, 2022), and includes work from our team (Hartshorne & Pustejovsky, 2021; Krishnaswamy, Pickard, Cates, Blanchard, & Pustejovsky, 2022; Pustejovsky & Krishnaswamy, 2019, 2022; Ghaffari & Krishnaswamy, 2022, 2023). Let us recall how we got here.

Computational Models in Cognitive Science

Computational models have been used in the study of complex systems across scientific disciplines –Physics, Biology, Chemistry, Geology, Medicine, Engineering, Economics, and, of course, Cognitive Science. They have been especially powerful in the study of nonlinear systems that are difficult to track with intuition or analytical thinking alone (Grubb, Moushegian, Heathcote, & Smith, 2020). Experiments are done by building different models, adjusting their variables, running the simulations, and observing the outcome. As technologies develop over time, the models provide more precise and accurate predictions and explanations of observed data.

In the case of Cognitive Science, computational modeling has been at the core from the start, with its founding claim that cognition is a form of information processing, coupled

with the emergence of the field of Artificial Intelligence (AI). First this came in the form of Symbolic AI (Turing, 1950; McCarthy, Minsky, Rochester, & Shannon, 1955; Newell & Simon, 1961), further defended in Philosophy by Chomksy (1959), Putnam (1967), and Fodor (1975), which later came into competition with Neural Network models, beginning with McCulloch and Pitts (1943) and Rosenblatt (1957). This lead to debates about the architecture of the human mind (Churchland, 1981; Fodor & Pylyshyn, 1988), continuing today to include Bayesian/Causal approaches (Pearl, 2000; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

However, the field of Cognitive Science defines itself as the study of the mind, which applies broadly not only to *actual* human and animal minds, but also potentially to alien minds, plant minds, silicon-based minds, and all other *possible* minds. What kinds of entities can be minds? Can there be 'pure' minds, as Chalmers (2023) claims, with no sensory connection to a world outside of it, or is sensory grounding required for having thoughts? What kinds of entities can a mind without language think about? Can minds like ours learn categories in a world with much fewer, or vastly different, regularities than the world we live in? Would we understand our world differently if our eyes were on our feet instead of our heads? These questions are relevant to a full understanding of the nature of the mental, what various minds can think about and the roles of sensation, action, and the environment.

Embodied and Embedded Cognition

The idea that the body and environment are crucial to cognition dates back arguably at least to Aristotle, but was also promoted in the last century by Husserl (1929), Merleau-Ponty (1962), and Heidegger (1975), and was recently revived (Gibson, 1966; Smith & Thelen, 1993; Hutchins, 1995; Clark, 1997; Lakoff & Johnson, 1999; Zahavi, 2005; Gallagher, 2006; Thompson, 2010). Roughly, the idea of *embodied* cognition is that accounts of cognition must include not only what happens inside the skull, but also what happens in an agent's sensory receptors and motor actuators. The idea of *embedded* cognition is that accounts must also include features of the world being sensed, represented, and acted upon. In the field of AI this pair of claims has come to be known as the symbol-grounding problem (Harnad, 1990), which was raised originally as a challenge to Symbolic AI, but holds for most Neural Network and Bayesian/Causal models as well.

On our more nuanced account, being causally connected with an external world is neither necessary nor sufficient for cognition. What we require is that the agent has sensorimotor representations that it *treats* as having arisen externally and that it tries to explain with a model of that external world. A brain in a vat, while in fact cut off from the world, can have meaningful thoughts as long as it meets this criterion. This does, however, rule out Large Language Models (LLMs), like OpenAI's GPT and even so-called 'Multimodal LLMs', like GPT-4V, LLaVA (Liu, Li, Wu, & Lee, 2024), and LLaVAR (Zhang et al., 2023). Some (e.g., Chalmers (2023)) argue that these systems are grounded because their linguistic data come from human users and/or their visual data come from cameras. But they don't *treat* their strings of text as utterances with communicative intent or their images as having been caused by anything outside of themselves. Their processing thus isn't *cognitive* because it isn't representational; it isn't aimed at being *about* anything (see Bender and Koller (2020) and Harnad (2024) for a more complete defense).

One approach to building embodied and embedded cognitive agents is to build physical, hardware sensors and/or motors to take information from the outside world and to act upon it— i.e., *robots*. While we feel that this is a step in the right direction, the field of Robotics is far from replicating human or animal vision, or any of our other sensory capacities, and it is far from creating life-like locomotion. Moreover, even as implementations of some possible minds, they are embedded only in our actual world. As we've been arguing, we need to experiment not only with how various possible minds might interact with the *actual* world; we want to explore how they would interact with the various *possible* worlds in which they might be embedded.

Virtual Agents in Virtual Worlds

This brings us to virtual robots, also known as *softbots*. Indeed, one way the field of Robotics has bypassed some of its engineering hurdles is by building virtual robots in virtual worlds. Engineers do this primarily to lower costs during the design of their physical robots. Because of this aim, they use virtual worlds that model the physics of our world, at least the aspects of our physics that are relevant to the functioning of their robots. The video game industry has recently made such real-world-like virtual environments available (see Collins, Chand, Vanderkop, and Howard (2021) for a review).

A growing number of researchers are using virtual embodied agents in virtual worlds for the study of cognition, particularly for modeling cognitive development in human toddlers. This includes the MIMo (Multimodal Infant Model) project being developed by Jochen Triesch's team (Mattern et al., 2022), which is an open-source platform that embeds a multimodal virtual toddler, with binocular vision, a vestibular system, proprioception, and touch perception, in a virtual world that uses the MuJoCo physics engine. A similar open source platform, VoxWorld, is developed and maintained by members of our team (Krishnaswamy et al., 2022), using VoxML (Visual Object Concept Modeling Language) to build on top of the Unity game engine a library of natural-kind objects and artifacts, with various shapes, sizes, surface properties, density, and afforded behaviors (Pustejovsky & Krishnaswamy, 2016, 2022), including *habitats* or configurations that condition such affordances (Pustejovsky, 2013). Because the platform is built on Unity, which visualizes objects from a player's point of view, it can be easily interfaced with a virtual agent that interacts with the objects in that world through its virtual sensors and movement actuators.

The VoxML platform allows researchers to design a range of possible agents to be embodied and embedded in various VoxWorlds. We can experiment with different 'innate' sen-

sory and motor abilities, learning algorithms, memory capacities, and innate theories of physics, biology, language, and other minds. Current extant agents include humanoid, simulated robotic, and self-exploring virtual toddler-like agents. These agents use the hybrid 'Best of All Worlds (BAW)' architecture being developed by our team to include the most promising elements of Symbolic, Neural-Network, and Embodied AI (Krishnaswamy et al., 2022; Hartshorne & Pustejovsky, 2021). The agents explore objects in their world, taking perceptual samples, and learning about objects' or events' intrinsic or extrinsic properties using various types of machine learning (Pustejovsky, Krishnaswamy, & Do, 2017; Krishnaswamy, 2017; Krishnaswamy & Pustejovsky, 2018; Ghaffari & Krishnaswamy, 2023).[1] We leave open whether such simulated minds constitute *synthetic* minds, or are *mere* simulations, analogous to simulated hurricanes, to be used in theorizing (see Searle (1980)).

The Unity-based VoxWorld platform also allows for the creation of a range of virtual worlds to be used in experimentation. VoxWorlds can be built with very different gravity from our world, or animals that are superficially similar to one another but have different dispositions, or even 'gruesome' worlds where objects change their perceptible properties at arbitrary times (Goodman, 1965). This flexibility allows researchers to explore questions about how a given type of mind (e.g., one like ours) might interact with worlds that are quite different from ours, what types of worlds they could learn in, understand, and talk about.[2]

Philosophical Puzzles about Representation

In this section, we recall three well-known philosophical puzzles about mental and linguistic representation. Later we will show how VW CogSci can help dissolve these puzzles by giving a view from the outside of entities in minds, entities in an environment, and relationships between them.

Even before the emergence of Cognitive Science, philosophers have puzzled over the nature, meanings, and acquisition of mental and linguistic representations. In ordinary parlance, we say things like, 'Abby hid when she heard a coyote because she *believed that coyotes are monsters* and she *desired that she stay safe*'. Several cognitive scientists have urged the elimination of such propositional attitudes (beliefs and desires) from scientific explanations of human behavior as they fail to map easily onto observable features of the brain and may be better understood as dispositions to behave than as explicit representations (Churchland, 1981; Stich, 1983; Dennett, 1989). Others have defended them as genuine mental entities, complex representations constructed in part from concepts, like COYOTE, MONSTER, and SAFE, as they explain reasoning, the systematicity and compositionality of thought, and symbolic language use (Fodor & Pylyshyn, 1988; Quilty-Dunn, Porot, & Mandelbaum, 2023).

But even among realists about propositional attitudes and concepts, puzzles persist when it comes to what we should count, e.g., as instances of *the belief that coyotes are monsters* or *the COYOTE concept*. Part of the problem, we suggest, is the assumption that such representations can be identified by their meanings or 'content' — i.e., by what they represent. What makes it the case that a given mental entity is a representation *of* the property/kind *coyote*? Internalists (Frege, 1948; Rosch, 1978; Segal, 2000; Prinz, 2002) hold that what makes a given mental entity a token of the COYOTE concept is its relation to other representations — of their furriness, four-legged-ness, distinct howl, and beliefs about their being wild, mammals, and so on. Externalists (Putnam, 1975; Kripke, 1980; Fodor, 1998; Burge, 2010), in contrast, hold that what makes something a token of the COYOTE concept is that it tends to be caused/activated by instances of coyotes in the outside world. Other theorists (Block, 1998; Chalmers, 2006), hold hybrid accounts, on which both internal and external factors are relevant. Others, still, hold that such concepts are constructed gradually over the course of development (Carey, 2009; Gopnik & Wellman, 2012), leaving the question of concept identity indeterminate.[3] Philosophers, mostly using the same old methods of armchair thought experiments that have been used for millennia, have created increasingly convoluted variations on these accounts (Prinz, 2002; Fodor, 1998; Laurence & Margolis, 1999), while others (e.g., Machery (2009)) take the accounts to be so convoluted that we should return to the complete elimination of beliefs and concepts from our science of the mind.

Next, we describe three philosophical puzzles that persist despite attempts to characterize concept and belief types. The puzzles are often dismissed as edge cases, but we will see later that similar cases are central to accounts of mental representation, particularly of their dynamical interactions with environments during the course of development.

Kripke's Case of Pierre in Londres

In Saul Kripke's (1979) *A Puzzle about Belief*, he describes the example of Pierre, who lives in France and hears about a beautiful city named 'Londres', which is the French name for London. Pierre tends to make statements like, 'Londres est jolie', which, as Kripke notes, we readily translate and ascribe to Pierre *the belief that London is pretty*. But as the thought experiment continues, Pierre moves to a run-down, dirty part of London, a city he is told is called 'London', and he comes to believe that this new town he lives in is ugly, not realizing that 'London' refers to the same city as 'Londres'.

[1] A philosophically curious feature of current versions is that they are dualistic in that the software running the agent's algorithms is outside of the virtual world the agent is 'embedded' in.

[2] It might be tempting to suppose that LLMs plausibly maintain something like a virtual world in this sense, however their world model would be internal to the agent, akin to a mental model, not an outside world in which the agent is embedded and trying to model. Moreover, it can be empirically demonstrated that they lack coherent world models. For example, Ghaffari and Krishnaswamy (2024) found that while they can correctly describe a ball or a coconut, they are unable to reason about the effect of a round object on the stability of a structure

[3] Indeed they leave concept types indeterminate in ways that align quite well with our position here.

Does Pierre now *believe that London is ugly*? Does he also still *believe that London is pretty*? His original representation hasn't changed, but we now want to retract the original ascription. Theorists twist and turn to accommodate this case, as it pulls on Externalist and Internalist intuitions, and it does so in opposing directions that are not eased by hybrid accounts. We will show later that this sort of case is not only common, but central to accounts of learning and development and is easily accommodated when we view the situation from the outside and talk about belief *tokens* instead of belief *types*.

Putnam's Case of Water on Twin Earth

Next, consider Hilary Putnam's (1975) Twin Earth thought experiment. Twin Earth is just like Earth in every superficial way, with a twin Florida, twins of Earth's mountains, animals, plants, people, etc. The only difference is that on Twin Earth the watery stuff that fills the oceans and lakes and nourishes its lifeforms is made of some other molecule, XYZ, instead of H_2O. As Putnam points out, it would be awkward to say that someone here on Earth and their twin on Twin Earth share *the concept WATER*, even if their internal mental entities are identical. When the earthling entertains her concept WATER, she is thinking about H_2O; when her twin entertains hers, she is thinking about some other substance, XYZ. Again, our Internalist intuitions pull us to say they have the same concept, but our Externalist ones make them distinct. Hybrid accounts don't ease the pain so much as describe it.

Putnam's Case of Jade

In the same paper, Putnam (1975) describes the case of jade, a set of stones with similar superficial appearances, but which in fact divide into two very different underlying mineral structures. Suppose we have someone, Peter, who doesn't know this. Can we, with our superior knowledge, ascribe to Peter *the belief that jade is green*? It is awkward to do so, at least without heavy elaboration. Consider Mary, a mineral scientist who is well-versed in the two kinds of stone, representing them with their corresponding scientific terms, 'jadeite' and 'nephrite'. Can we say of Mary that she has *the belief that jade is green*? As for the English word, 'jade', what is its meaning? Does it fail to have a meaning since using it presupposes, falsely, that there is a uniform mineral kind that it picks out? Or does it represent a disjunctive category, *jadeite or nephrite*, unbeknownst to most users? Again, our intuitions pull us in multiple directions.

On the flip side, it has come to our attention that rubies and sapphires, while different in color, are in fact the same mineral, known as *corundum*. It is only because of impurities, traces of chromium versus iron and titanium, that they reflect light to give their reddish or bluish appearance (Ward, 2003). Knowing this, can we ascribe to each other *the belief that sapphires are blue*?

Beliefs and Concepts with VW CogSci

As Timothy Williamson (2007) argued in his influential book, a major problem with thought experiments is that they are coarse descriptions with a lot of details left to be filled out by our own imaginations, often produced by our prior theories. In Kripke's case of Pierre, we assume initially that he has only one representation for the city of London. Later in the scenario, this no longer holds, and our belief ascriptions change. One key advantage of VW CogSci for thought experiments is that it forces a more complete and explicit fleshing-out of the scenarios under consideration. Next, we will see how the method can help resolve philosophical puzzles like the ones described in the previous section. Moreover, we will see how the method allows us to track the development of mental representations as agents gain experience with entities in an environment and with other agents who share that environment, much in the spirit of the Constructivist approach described by Gopnik and Wellman (2012).

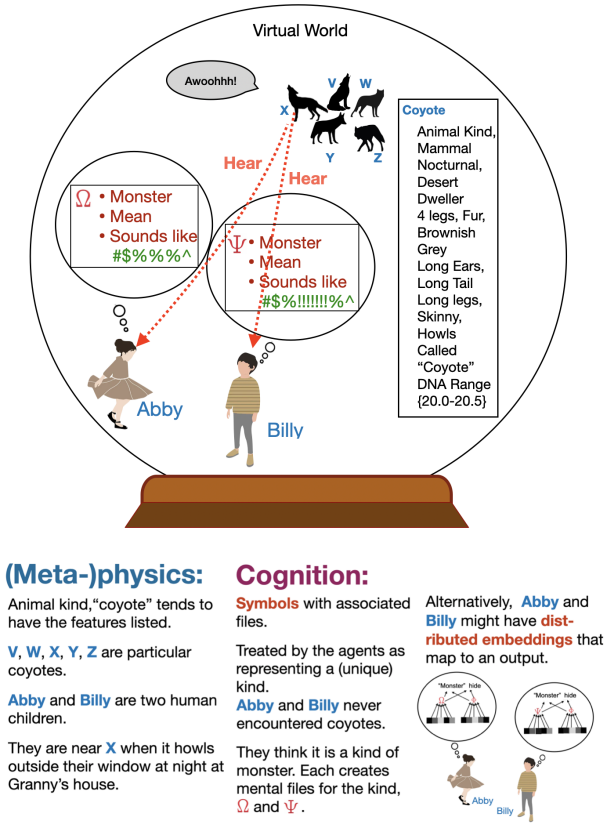

Figure 1: A virtual world in which Abby and Billy hear coyotes, and each form a concept token for the kind.

Let's suppose we build a virtual world that we furnish the with instances of various animal kinds — dog, wolf, coyote, cat, tiger, such that, e.g., any entity that is a coyote is likely to be furry, tends to howl, has lungs, and so on (see Figure 1). Such essentialism, with objective facts about whether an entity is a coyote, may or may not be in the metaphysics of our world, but the point is that we can experiment with this too.

Next, suppose we build two virtual human toddlers, Abby and Billy, and in the simulation they hear coyotes outside Granny's house at night. The coyotes produce virtual sound

waves that cause auditory sensations in the children. Suppose they both infer that they are hearing some new kind of entity, and they each create a mental file for this newly detected kind, label the kind Ω and Ψ respectively, and store what they believe to be the likely features of the kind.[4] (Alternatively, we could build Abby and Billy with a Neural Network architecture, in which case they might form distributed embeddings instead of symbols and files, as shown in the figure.)

Already at this point, theorists might begin to argue over whether Abby and Billy have *the COYOTE concept*, or *the belief that coyotes are monsters*, but it's not clear what that adds to our understanding. VW CogSci gives us a full view from the outside, so we can simply describe Abby's and Billy's token labeled files, Ω and Ψ (or their token distributed embeddings), run the simulation, and observe their attempts to coordinate with other agents and the entities in their world. Abby and Billy have very little information about coyotes, and some of it is false, yet they have representations that allow them to think about what in fact are coyotes, and add knowledge about coyotes as they encounter more of them.

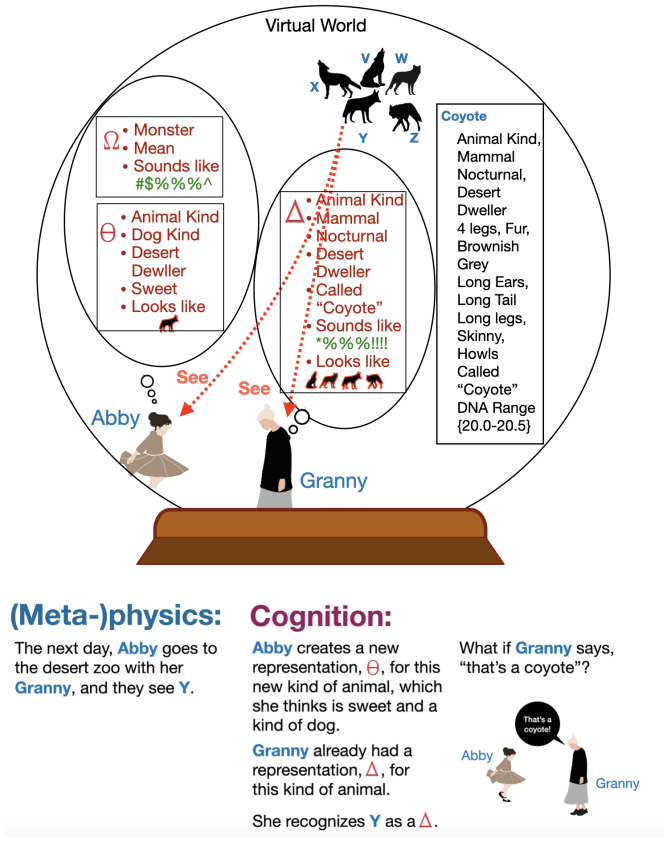

Figure 2: Granny takes Abby to the desert zoo and they see coyotes. Abby creates a new file and label for what she thinks is a new kind of entity.

Next, suppose that Granny takes Abby to the desert zoo the following day, and they see some skinny-legged, furry, dog-like animals which in fact are coyotes. Not knowing these are the same kind of entity as what she heard the night before, Abby creates a new file, θ, (or distributed network) for this category (see Figure 2). Notice that Abby is a lot like Pierre in Kripke's (1979) example; she has two representations that, unbeknownst to her, refer to the same thing. Granny already had encountered coyotes many times before so she already had a labeled file for coyotes, Δ, and recognized the instances.

Do Abby and Granny share *the concept COYOTE*? Does Abby have *the belief that coyotes are mean monsters* or *the belief that coyotes are sweet dogs*? Again, the answers aren't obvious. Suppose Granny tells Abby that these animals are called 'coyotes'? Would she then know *what coyotes are*? Would she know *the meaning of the word 'coyote'*? Abby would presumably still have two mental files which she treats as representing two different kinds of entity. We don't have English words for her two categories, so we aren't able to distinguish them with ordinary language. Again, our stance is that it isn't helpful to try to settle such matters. With the whole picture from the outside, we can simply consider their respective concept and belief *tokens* and observe how easily they coordinate with each other and the entities in their world.

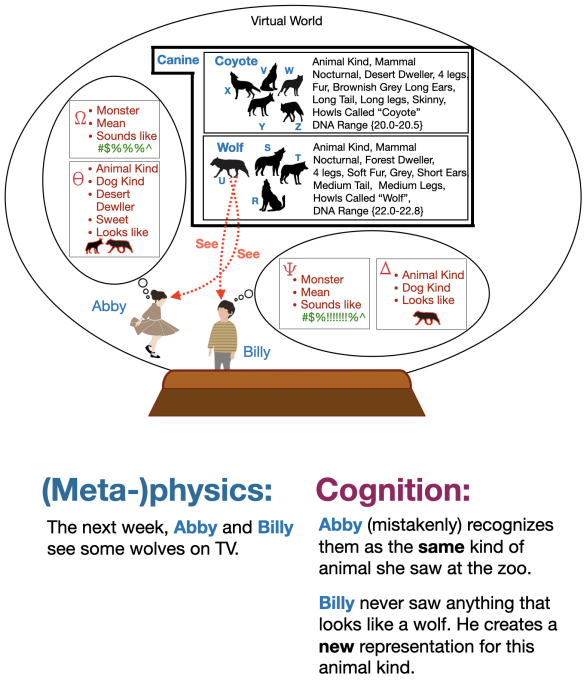

Figure 3: Abby and Billy see wolves on TV. Abby adds details to one of her files while Billy creates a new one.

We can make matters worse by supposing that Abby and Billy later see what in fact are wolves on TV (Figure 3). Suppose Abby thinks they look like the coyotes she saw at the zoo, so she adds them to her θ file. Billy has never encountered this appearance, so he creates a new file, labeling it Δ. Now Abby's θ is a lot like JADE in Putnam's (1975) example; she treats as one animal-kind what in fact are two kinds.

[4]This variation of the labeled-files view of concepts follows the Baptism model proposed by Oved (2015).

Do either Abby or Billy have *the belief that wolves are furry*? Do they share *the concept WOLF*? Is there any sense in which Billy and Granny *share* a concept, given they both have files they label Δ? Again, with the full picture from the outside, it's not helpful to try to answer these questions. Abby and Billy are children simply trying to learn about their world, carving its joints as best they can as they go along. At this stage in their learning, their carvings fail to correspond to the objective joints in their world, but as we run the simulation and they continue to gain more experience, we will be able to observe whether their models become more aligned with their world. Hopefully Granny will also continue to learn as she ages. She might discover that what people call 'coyotes' in her world in fact divide into two different species that cannot mate and have deep biological distinctions. We need room for such development in our theory of beliefs and concepts. These mistakes and revisions are expected, even healthy, and we can fully describe such cognitive development by appeal to the agents' respective concept and belief tokens.

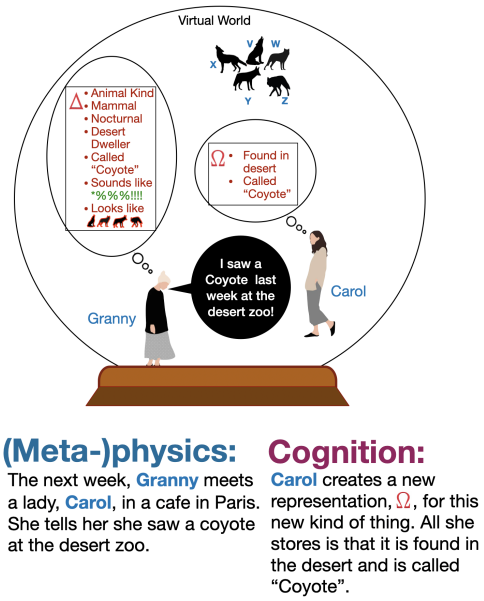

Figure 4: Granny says to Carol, "I saw a coyote at the desert zoo". Carol creates a file for the kind.

Finally, suppose that Granny goes to Paris the next week and meets a woman named Carol. Excited to share about her life in the desert, Granny says to Carol, "I saw a coyote last week at the desert zoo" (see Figure 4). Suppose that Carol infers that 'coyote' is the name for some kind of thing that is found in the desert. She creates a file, labels it Ω, and stores the minimal information she has. [5] Does Carol have *the COYOTE concept*? She is able to think about what in fact are coyotes, and she can now learn more about them, by asking, e.g., "Are coyotes animals?", or pointing at something and asking "Is that a coyote?". But as for *the COYOTE concept*, there is no added value in deciding whether she has it. Notice that if we replace the word 'coyote' in this example with an *empty name*, like 'witch', 'ghost', or 'vampire', we would have no trouble describing the representations as tokens.

The scenarios described above are typical of the learning process for young children as they encounter new entities in their world. VW CogSci allows us to step outside both the mind and the world in question, so we can eliminate problematic talk about belief and concept *types*, while keeping their tokens. We can then run simulations of Abby and Billy interacting with the entities in their world, and observe how their representations shift, merge, split, how well they correspond to their worlds, and how easily the agents interact with one another through language and gesture. In hindsight, perhaps the real puzzle is why philosophers have been twisting themselves into pretzels for millennia to describe mismatches between an agent's carving of a world and an objective one.

Conclusions

This paper gave philosophical motivations for a method we call *Virtual World Cognitive Science* (VW CogSci), in which researchers use virtual embodied agents that are embedded in virtual worlds to explore questions in the field of Cognitive Science. It then showed how the method can be used to dissolve many philosophical puzzles about mental and linguistic representation and test complex theories about the development of such representations.

After describing the method of VW CogSci, we defended it on the basis of (1) the virtues of computational modeling in the articulation and testing of complex theories in science; (2) the view that mental and linguistic representations are best understood in part by appeal to an agent's sensori-motor interactions with its (assumed) environment; and (3) the claim that the science of the mind must go beyond actual human and animal minds in our world, to include accounts of what kinds of minds are possible and in what kinds of worlds. Cognitive Science, just as *any* science, seeks models that posit a set of entities and regularities that explain our observations, support counterfactuals, and can be tested by interventions. A full model of the mental will thus include relationships between variables in minds, bodies, and worlds.

We then turned to mental and linguistic representations and showed how VW CogSci adds rigor to their study. First, we showed that by taking a god's-eye view, the method eliminates the need for problematic talk of belief and concept *types*, such as *the belief that cats are silly*, and *the concept CAT*, while preserving the explanatory power of belief and concept *tokens* in individual cognizers' minds. Second, we showed how the method allows for the study of various possible minds in various possible worlds to explore questions about the nature, meaning, and development of mental representations by playing out their complex interactions in a simulation rather than trying to track them by armchair analysis.

[5] This situation is similar to the one described by Putnam (1975) for the English terms 'Elm' and 'Beech', for which most people have very few associated representations, besides the belief that they name two different kinds of tree.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) on grant IIS 2033938 to Boston College and grant IIS 2033932 to Brandeis University, and by the U.S. Army Research Office (ARO) on grant W911NF-23-1-0031 to Colorado State University. The views expressed herein do not reflect the official position of the U.S. Government. We'd also like to thank our anonymous reviewers as well as David Barner, Gedeon Deák, Ian Fasel, Gail Heyman, Terry Horgan, Mengguo Jing, Joshua Knobe, Carlos Montemayor, Seth Neiman, Shaun Nichols, and Matthew Stone for helpful discussion and comments on earlier drafts. Any errors or omissions are the responsibilities of the authors.

References

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Yoshikawa, Y., Ogino, M., & Yoshida, C. (2009, 06). Cognitive developmental robotics: A survey. *Autonomous Mental Development, IEEE Transactions on*, *1*, 12 - 34. doi: 10.1109/TAMD.2009.2021702

Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185–5198). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.463 doi: 10.18653/v1/2020.acl-main.463

Block, N. (1998). Conceptual role semantics. In E. Craig (Ed.), *The routledge encyclopedia of philosophy* (pp. 242–256). Routledge.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

Burge, T. (2010). *Origins of objectivity*. Oxford, GB: Oxford University Press.

Cangelosi, A., & Schlesinger, M. (2015). *Developmental robotics: From babies to robots*. MIT press.

Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.

Chalmers, D. J. (2006). Two-dimensional semantics. In E. Lepore & B. Smith (Eds.), *The oxford handbook to the philosophy of language.* Oxford University Press.

Chalmers, D. J. (2023). Does thought require sensory grounding? from pure thinkers to large language models. *Proceedings and Addresses of the American Philosophical Association*, *97*.

Chomksy, N. (1959). Review of skinner's verbal behaviour. *Language*, *35*, 26–58.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, *78*(2), 67–90.

Clark, A. (1997). The dynamical challenge. *Cognitive science*, *21*(4), 461–481.

Collins, J., Chand, S., Vanderkop, A., & Howard, D. (2021). A review of physics simulators for robotic applications. *IEEE Access*, *9*, Article number: 9386154 51416–51431.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Fodor, J. A. (1975). *The language of thought.* New York: Thomas Y. Crowell.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford, GB: Oxford University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Frege, G. (1948). Sense and reference. *Philosophical Review*, *57*(3), 209–230. doi: 10.2307/2181485

Gallagher, S. (2006). *How the body shapes the mind*. Clarendon Press.

Ghaffari, S., & Krishnaswamy, N. (2022). Detecting and accommodating novel types and concepts in an embodied simulation environment. In *Annual conference on advances in cognitive systems (acs). cognitive systems foundation.*

Ghaffari, S., & Krishnaswamy, N. (2023). Grounding and distinguishing conceptual vocabulary through similarity learning in embodied simulations. *IWCS 2023*, *305*, 305.

Ghaffari, S., & Krishnaswamy, N. (2024). Exploring failure cases in multimodal reasoning about physical dynamics. *arXiv preprint arXiv:2402.15654*.

Gibson, J. J. (1966). The senses considered as perceptual systems.

Goodman, N. (1965). *Fact, fiction, and forecast*. Cambridge, Mass.: Harvard University Press.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085-1108.

Grubb, A. L., Moushegian, A., Heathcote, D. J., & Smith, M. J. (2020). Physics and computational modeling of nonlinear transverse gust encounters. In *Aiaa scitech 2020 forum* (p. 0080).

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346.

Harnad, S. (2024). *Language writ large: Llms, chatgpt, grounding, meaning and understanding.*

Hartshorne, J., & Pustejovsky, J. (2021). *A playground and proposal for growing an artificial general intelligence.* (Tech. Rep.). Boston College and Brandeis University.

Heidegger, M. (1975). *The basic problems of phenomenology* (Vol. 478). Indiana University Press.

Husserl, E. (1929). *Cartesian meditations: An introduction to phenomenology*. Springer Science & Business Media.

Hutchins, E. (1995). *Cognition in the wild*. MIT press.

Kripke, S. A. (1979). A puzzle about belief. In *Meaning and use: Papers presented at the second jerusalem philosophical encounter april 1976* (pp. 239–283). Springer.

Kripke, S. A. (1980). *Naming and necessity: Lectures given to the princeton university philosophy colloquium*

(D. Byrne & M. Kölbel, Eds.). Cambridge, MA: Harvard University Press.

Krishnaswamy, N. (2017). *Monte carlo simulation generation through operationalization of spatial primitives*. Brandeis University.

Krishnaswamy, N., Pickard, W., Cates, B., Blanchard, N., & Pustejovsky, J. (2022). The voxworld platform for multimodal embodied agents. In *Proceedings of the thirteenth language resources and evaluation conference* (p. 1529-1541).

Krishnaswamy, N., & Pustejovsky, J. (2018). An evaluation framework for multimodal interaction. In *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018)*.

Lakoff, G., & Johnson, M. (1999). Philosophy in the flesh—the embodied mind and its challenge to western thought.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (pp. 3–81). MIT Press.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *Advances in neural information processing systems*, *36*.

Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: a survey. *Connection science*, *15*(4), 151–190.

Machery, E. (2009). *Doing without concepts*. New York: Oxford University Press.

Mattern, D., López, F. M., Ernst, M. R., Aubret, A., & Triesch, J. (2022). Mimo: A multi-modal infant model for studying cognitive development in humans and ais. In *2022 ieee international conference on development and learning (icdl)* (p. 23-29). doi: 10.1109/ICDL53763.2022.9962192

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, *27*(4), 12–12.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*, 115–133.

Merleau-Ponty, M. (1962). *Phenomenology of perception* (D. A. Landes, Ed.). New York: Routledge.

Newell, A., & Simon, H. A. (1961). Gps, a program that simulates human thought.

Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, *11*(2), 265–286.

Oved, I. (2015). Hypothesis formation and testing in the acquisition of representationally simple concepts. *Philosophical Studies*, *172*(1), 227–247. doi: 10.1007/s11098-014-0291-2

Pearl, J. (2000). Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*, *19*(2), 3.

Prinz, J. J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. MIT Press.

Pustejovsky, J. (2013). Dynamic event structure and habitat theory. In *Proceedings of the 6th international conference on generative approaches to the lexicon (gl2013)* (pp. 1–10).

Pustejovsky, J., & Krishnaswamy, N. (2016, May). VoxML: A visualization modeling language. In N. Calzolari et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 4606–4613). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from `https://aclanthology.org/L16-1730`

Pustejovsky, J., & Krishnaswamy, N. (2019). Situational grounding within multimodal simulations. *arXiv preprint arXiv:1902.01886*.

Pustejovsky, J., & Krishnaswamy, N. (2022). Multimodal semantics for affordances and actions. In *Human-computer interaction. theoretical approaches and design methods: Thematic area, hci 2022, held as part of the 24th hci international conference, hcii 2022, virtual event, proceedings, part i* (p. 137-160). Springer International Publishing.

Pustejovsky, J., Krishnaswamy, N., & Do, T. (2017). Object embodiment in a multimodal simulation. In *Aaai spring symposium: Interactive multisensory object perception for embodied agents.*

Putnam, H. (1967). Psychological predicates. In W. Capitan & D. Merrill (Eds.), *Art, mind, and religion* (p. 37-48). University of Pittsburgh Press.

Putnam, H. (1975). The meaning of "meaning".

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*, e261. doi: 10.1017/S0140525X22002849

Rosch, E. (1978). Principles of categorization. In A. Collins & E. E. Smith (Eds.), *Readings in cognitive science, a perspective from psychology and artificial intelligence* (pp. 312–22). Morgan Kaufmann Publishers.

Rosenblatt, F. (1957, January). *The perceptron - a perceiving and recognizing automaton* (Tech. Rep. No. 85-460-1). Ithaca, New York: Cornell Aeronautical Laboratory.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–57. doi: 10.1017/s0140525x00005756

Segal, G. (2000). *A slim book about narrow content*. MIT Press.

Smith, L. B., & Thelen, E. (1993). *A dynamic systems approach to development*. CogNet.

Stich, S. P. (1983). *From folk psychology to cognitive science: The case against belief.* the MIT press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Thompson, E. (2010). *Mind in life*. Cambridge, MA: Harvard University Press.

Turing, A. M. (1950). *Computing machinery and intelligence*. Springer.

Ward, F. (2003). *Rubies & sapphires*. Gem Book Publishers.

Williamson, T. (2007). *The philosophy of philosophy*. John Wiley & Sons.

Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective*. MIT press.

Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., & Sun, T. (2023). Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.