# Virtually Impossible: Obstacles to Generalizing between Simulated and Real Humans

**Sean Kugele[1*], Zachariah A. Neemeh[2], Christian Kronsted[3], Shaun Gallagher[4], Stan Franklin[†]**

[1]Department of Computer Science, Rhodes College, Memphis, TN, USA

[2]Stanford Online High School, Stanford University, Stanford, CA, USA

[3]Department of Philosophy, Merrimack College, North Andover, MA, USA

[4]Department of Philosophy, University of Memphis, Memphis, TN, USA

[†]Deceased, January 2023

*** Correspondence:**
Corresponding Author
seankugele@gmail.com

**Keywords: cognitive modeling, autonomous agents, embodied cognition, culture, synthetic approaches, animat approach, cognitive architectures, LIDA.**

**Abstract**

The validity of a virtual human-based research methodology, in which simulated humans are used to generate knowledge about real humans, depends on substantiating multiple correspondence claims which are *currently indefensible*. One must substantiate that real and virtual humans are sufficiently similar with respect to their (1) control structures, (2) environments and embodied experiences, (3) adaptive histories and attunements, (4) social and cultural contexts, and (5) institutional contexts. If one's confidence in any of these correspondences is undermined, then the foundation of this approach will crumble.

Unfortunately, technological limitations and our fragmentary understanding of minds will severely constrain the similarities between real and virtual humans for the foreseeable future. As a result, attempts to generalize empirical findings from virtual humans to real humans will prove ill-founded, and are likely to fail. Therefore, we believe that alternative research methodologies that focus on understanding mechanisms of mind more broadly, and cultivate the gradual acquisition of enabling technologies and engineering competences, are needed in the interim. We describe two such alternative approaches here, and speculate on their usefulness and viability in practice.

## 1 Introduction

The virtual human research methodology is exemplified by DiPaola et al. (2021) in a research topic they proposed for *Frontiers in Psychology*. They stated,

> This research topic centers on the methodology of understanding systems by building them, specifically the construction of autonomous computer-generated humans as a research methodology. It is directly supported by the dramatic increase in the graphical quality of computer-generated humans: virtual humans appear indistinguishable from real humans, providing a unique opportunity to push more realistic cognitive and behavioral models… In building models that drive artificial humans, we are asking questions relevant to the understanding of the human mind…. the emphasis of this [methodology] is on the use of virtual humans to embody models and testing them in real-time interaction. The cornerstone is that the model's quality is assessed by the quality of the interaction between the virtual human, controlled by the model, and the biological human.

Researchers applying this methodology observe and manipulate virtual humans in simulated environments to support or challenge cognitive theories, and to generate new hypotheses about real humans. As with the use of natural animal models, this approach's validity fundamentally depends on the degree of correspondence between model and target species, that is, how human-like these virtual humans really are.

Traditional animal-model approaches are predicated on millions of years of shared evolutionary heritage and the assumption that the resulting genetic, metabolic, developmental, or behavioral systems are substantially conserved between our species and theirs. For example, advocates for the use of animal models such as mice are quick to point out the many similarities between mouse and human genomes (*Why Are Mice Considered Excellent Models for Humans?*, n.d.). Despite this, generalization errors between these animal models and humans can, and often do, occur. For example, less than 8% of promising cancer treatments developed in natural animal models have led to successful medical interventions in humans, and in one particularly notable case, a promising cancer drug caused catastrophic organ failure in humans with doses *five hundred times lower* than those safe in non-human animal studies (Mak et al., 2014). A similarly low success rate has been observed in the development of treatments for central nervous system disorders (such as Alzheimer's and schizophrenia) based on non-human animal models (Geerts, 2009).

Within the domain of human cognition and behavior, correspondence problems can be particularly acute, since social and cultural factors can play significant roles. While traditional views of culture in social psychology and cognitive science have taken it to be an external force influencing cognition (e.g., Hofstede, 2001), recent research in cultural neuroscience (Hanakawa et al., 2003; Kitayama & Park, 2010; Seligman et al., 2016), cognitive anthropology and archaeology (Henrich, 2016; Overmann, 2017; Overmann & Wynn, 2019), and enactive cognition (Gallagher, 2013; Hutto et al., 2020; Petracca & Gallagher, 2020) demonstrate that culture pervasively modulates cognition and brain processes. Institutions, practices, technologies, and other people act as external scaffolds for many cognitive processes. For example, in a legal context, a judge relies on institutional norms and practices, codified laws, legal precedents, social expectations, and interactions with other agents with well-defined roles (e.g., jury members, defendants, and prosecutors). This external scaffold constitutively enables the legal judgment that the judge makes; it is not the sole achievement of the judge's brain. Similarly, scientists make discoveries in collaborative laboratories (Slaby & Gallagher, 2014), and economists (as well as consumers and producers) make decisions within financial markets (Gallagher et al., 2019; Petracca & Gallagher, 2020). These social and cultural factors can be challenging to account for in experimental settings. Yet neglecting to include them can undermine the significance of our experiments, and render our models ineffectual as scientific tools.

If establishing a correspondence between natural animal models and humans is difficult, how much more challenging will it be to justify similarities between software agents and humans? There seems little reason to expect that the minds of these engineered beings are any more human-like than non-human animals, and, on the contrary, we have every reason to suspect that they will share fewer similarities with us than we do with our biological cousins (e.g., other mammals, reptiles, or even insects). We do not share an evolutionary heritage, bodily substrate, environment, or social, cultural, and institutional contexts with artificial humans.

*We will argue that the validity of the virtual human methodology depends on substantiating multiple correspondence claims which are currently indefensible.* One must substantiate that (1) virtual humans are autonomous agents with "control structures" (Newell, 1973) that are sufficiently similar to real humans; (2) their virtual environments, and interactions with those environments, are sufficiently similar to those of humans in the physical world; (3) their experiential and evolutionary histories result in sufficiently human-like adaptations and attunements; (4) social and cultural contexts within their virtual environments afford human-like opportunities for interactions with other virtual humans and their virtual world; and (5) institutional contexts, including norms, practices, and related technologies, are available in the virtual environment to externally scaffold the activities of virtual humans.

In these arguments, we take for granted that cognition is fundamentally dependent on body and environment—a core tenet of embodied cognition. Unlike classical cognitivism, which views minds as abstract information processors analogous to computers, embodied cognition holds that an agent's mind is inseparable from its sensorimotor engagements with the world (M. Wilson, 2002). From this perspective, movements, affects, motivations, and social interactions are the primary driving forces of an agent's mind. Furthermore, natural (biological) agents are situated within, and a part of, ecological niches, and their cognitive capabilities develop in service of actions within those niches (Franklin, 1995, Chapter 16; Varela et al., 1991/2016).

While it may be *theoretically* possible to engineer virtual humans and simulated environments that satisfy the five correspondence conditions mentioned above, our current technological limitations and fragmentary understanding of minds will severely constrain the obtainable correspondence between real and virtual humans for the foreseeable future. Consequently, we believe the virtual human methodology is currently untenable, and that we should consider other, more tractable, options in the interim.

After establishing our core arguments in Section 2, we explore two alternative approaches to understanding minds in Section 3 that sidestep the aforementioned correspondence problems. These approaches are *bottom-up* and *incremental* synthetic approaches (Franklin, 1995, pp. 9–10) that replace the *resemblance-based* evaluation criterion used in the virtual human research methodology with a *performance-based* criterion that judges software agents based on their ability to produce adaptive behaviors in naturalistic virtual ecological niches. While our discussion of the feasibility and usefulness of these alternative approaches is largely speculative, they, nevertheless, provide an instructive contrast with the virtual human methodology.

## 2 The Virtual Human Methodology and Its Correspondence Problems

The virtual human methodology is analogous to the use of natural animal models as experimental proxies for real humans, but with species of engineered, artificial minds as the proxies. A prerequisite of such approaches is that researchers must substantiate that a correspondence exists between a model species (e.g., virtual humans) and the target species (e.g., real humans). If one's faith in this correspondence is undermined, then the foundation of the approach crumbles.

In the remainder of Section 2, we will argue that establishing a correspondence between virtual and real humans is not yet possible. This is due to multiple issues that weaken the *external validity*[1] of the proposed virtual human research methodology. While some of these issues are common to all animal modeling approaches, others stem from (or are exacerbated by) the engineered reality that this research paradigm requires. In particular, this approach forces researchers to either simulate *all aspects* of the real world, or defend claims about the irrelevance of those things they have neglected to include. Since the former is out of the question, a researcher's only recourse is to argue for the sufficiency of their impoverished renderings of humans and the real world. In particular, they must substantiate claims that real and virtual humans are sufficiently similar with respect to their control structures (Section 2.1), environments and embodied experiences (i.e., *Umwelten*) (Section 2.2), adaptive (personal and evolutionary) histories and attunements (Section 2.3), social and cultural contexts (Section 2.4), and institutional contexts (Section 2.5). Within each section, we will contend that various aspects of virtual humans and their environments are infeasible to simulate due to technological and theoretical limitations. While these aspects are deeply intertwined, we separate them here for the sake of clarity.

## 2.1 Correspondence of Control Structures

We define minds, both natural and artificial, as *control structures for autonomous agents* (see Franklin, 1995). While Newell (1973) explained the idea of a "control structure" through a computer programming analogy, control structures can be more broadly defined as those mechanisms that enable autonomous agents to answer the question, "What do I do next?" We follow Franklin and Graesser's (1997) definition of an autonomous agent as "*a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future*" (Franklin & Graesser, 1997, p. 25). According to this definition, autonomous (software) agents differentiate themselves from non-agential "programs" by their situated and embedded relationship with an environment, and their selection of actions that further their *own* agenda.

Given two minds, for example, a virtual human and a real human mind, ceteris paribus, one might try to establish a correspondence between their control structures by simply comparing the observable behaviors they produce. However, we contend that this purely behavioral approach is insufficient.

In 1950, Turing (1950) proposed his famous "imitation game" as a standard by which one could answer the question, "Can machines think?" It is based on the idea that if a human interrogator cannot tell the difference between a human's and a machine's behaviors (in particular, their responses to the interrogator's questions), then we should attribute to the machine a capacity for thought. The imitation game operationalizes a notion of thought and intelligent behavior that is *agnostic* of its underlying causes (i.e., the human's and machine's control structures). For Turing's purposes, this was completely adequate. However, some have misinterpreted Turing as implying that a machine passing such a test *thinks like a human*. To the contrary, Turing (1950) wrote,



[1] External validity is defined as "the extent to which research findings derived in one setting, population or species can be reliably applied to other settings, populations and species" (Pound & Ritskes-Hoitinga, 2018, p. 2).

> May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection. (Turing, 1950, p. 435)

While this objection may be irrelevant for establishing that a machine *thinks*, it cannot be ignored if one's purpose is to establish a correspondence between artificial and human minds. A human interrogator may be convinced that two minds produce similar behaviors, but that conviction is not sufficient to conclude that those behaviors originate from similar control structures.

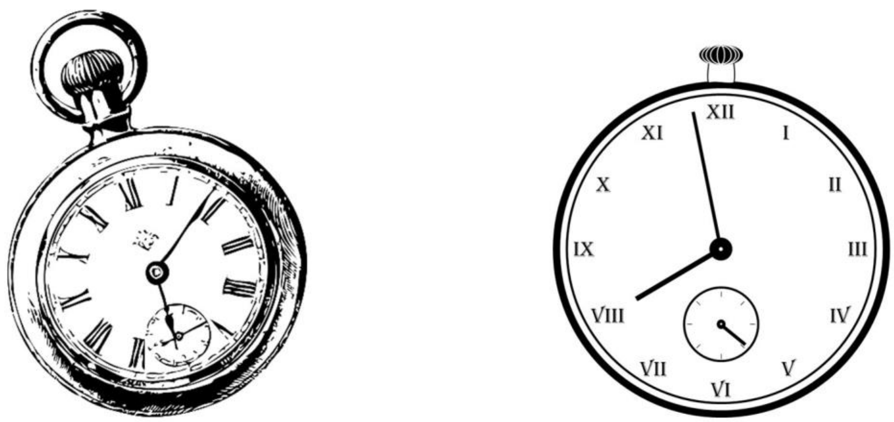

**Figure 1. A stem-wind, stem-set pocket watch (left) and its digital doppelganger (right).**

As a simple thought experiment, consider a 19th-century "stem-wind, stem-set" pocket watch (see Figure 1, Left). Even if we knew nothing about analog clocks or the mechanics of 19th-century pocket watches, we could easily discern (after a few hours of observation) that each of its three "hands" appear to rotate at predictable rates. The smallest hand rotates approximately six degrees *per second*. The long, thin hand rotates approximately six degrees *per minute*. And the medium-length, thick hand rotates approximately 30 degrees *per hour*. Armed with this knowledge and a few lines of code, we could produce a seemingly perfect digital doppelganger (see Figure 1, Right) of the real pocket watch. However, the underlying control structures are wholly dissimilar. As a result of these differences, our virtual watch fails us in almost every way as a model of the original watch. It fails to support the generation of new hypotheses and predictions about the real watch (e.g., what happens to its hands' rotation rates when the watch's hairspring tension is increased?). It fails to consider real-world operating conditions (i.e., the physical context) that can directly affect its observable behaviors. These include the effect of ambient temperature on its precision, the gradual cumulative effects of friction on its accuracy, or the potentially catastrophic effects of a strong magnetic field or emersion in water on its operations. And it fails to simulate any behaviors of the real watch that it was not explicitly programmed to mimic (e.g., the ticking sounds emitted by the real watch, or the way that its hands inexplicable stop rotating after several days unless the watch is wound). Most importantly, it fails to help us comprehend how the real pocket watch *works* (i.e., what physical forces and principles govern the movement of its hands?). And is that not the point of this whole endeavor?

In the 70 years since Turing proposed the imitation game, the script has changed dramatically. Thinking machines have been created that match or surpass the best efforts of skilled humans in contests that few would have believed possible a few decades ago. Machines have beaten the world's greatest chess (Campbell et al., 2002; Silver, Hubert, et al., 2017) and Go players (Silver, Schrittwieser, et al., 2017), Jeopardy champions (Ferrucci et al., 2010), and e-sports professionals (Vinyals et al., 2019). They have outperformed trained medical professionals at detecting lung cancer in diagnostic images (Ardila et al., 2019) and human experts on some language comprehension tasks (Devlin et al., 2018). We take for granted our AI-based digital assistants (like Alexa, Google Assistant, and Siri), using them habitually as cognitive supports, and even talking to them as if they were humans. And we appear to be on the verge of an era of self-driving cars and other autonomous vehicles. Every time we draw a line in the sand and say, "machines will never do this," we are invariably wrong. While there are currently no software agents that can reliably "win" Turing's imitation game, we regard this milestone as inevitable. *When they do*, it is critical that we understand the context and purpose of that original challenge, and not make the mistake of assuming that similar behavior necessitates similar minds.

This is particularly critical today, as recent computational and algorithmic advances have made it possible to use massive amounts of data to train models (e.g., "deep" neural networks) that *mimic* how humans behave in various contexts. For example, ***B**idirectional **E**ncoder **R**epresentations from **T**ransformers* (BERT; Devlin et al., 2018) is an artificial neural network (ANN) architecture that *outperformed* human "experts" on several language comprehension and production tasks, but, in spite of this, few would claim that BERT thinks or learns like a human. More recent network architectures like ***G**enerative **P**re-trained **T**ransformer **3*** (GPT-3; Brown et al., 2020) have further raised the bar on what can be achieved with extremely massive amounts of data and equally massive models[2]. From an engineering perspective, technologies like BERT and GPT-3 are marvels that will likely result in many useful tools for humanity. However, as cognitive scientists, we need to guard against making incorrect assumptions about the minds of these tools. They learn, and almost certainly think, in distinctly unhuman-like ways, even though their behaviors may suggest otherwise.

## 2.2 Environments and *Umwelten*

Cognition is not an isolated or purely internal process that is hidden away inside of agents; it is the product of agents being embedded in and coevolving with their environments. Simon (1996) illustrated this interplay by considering the trajectory of an ant traveling on a beach. When examined in isolation, the ant's behaviors look exceedingly complex. However, when we realize that the ant's path merely reflects the surface of the beach, the source of the complex trajectory becomes clear. The ant is *coupled* to its environment, and it is only in examining them *together* that its behaviors begin to make sense. This also applies to human behavior.

Given that environmental changes can produce behavioral changes in agents, one might wonder how realistic virtual environments must be for human-like behaviors to emerge. Afzal et al. (2020) recently surveyed 82 roboticists and found that many believed there was a significant "reality gap" between today's simulators and the real world. Participants complained that "simulation can produce unrealistic behaviors that would not occur in the real world" (Afzal et al., 2020, p. 3), and that accounting for all relevant physical phenomena can be challenging. Some perceived that this reality gap was large enough that they regarded simulation as *infeasible* for testing their robots. This



[2] GPT-3 has 175 billion model parameters, and was trained on half-a-trillion encoded linguistic tokens. It received a great deal of media attention because it has been claimed that the *synthetic* news articles generated by the model are practically indistinguishable from *real* news articles. For example, Brown et al. (2020) noted, "mean human accuracy at detecting articles that were produced by the [GPT-3] 175B parameter model was barely above chance."

suggests that we may currently lack the engineering knowledge and technical "know-how" to create realistic simulators for robots, let alone virtual humans.

Humans have a tremendous variety of sensors (e.g., visual, auditory, tactile, gustatory, olfactory, proprioceptive, nociceptive, and thermoreceptive). Unfortunately, today's simulated environments are almost exclusively focused on visual modalities. While vision likely dominates our sensations, an expansive range of other sensory phenomena shape our richly *multi-modal* perceptual experiences, and synergistically combine to produce our perceptual *Umwelten* (Uexküll, 2010). For example, humans have four different mechanoreceptors in the skin (Merkel receptors, Meissner corpuscles, Ruffini cylinders, and Pacinian corpuscles), each responding to different kinds of pressure and stretching stimuli. Together they lead to the holistic perception of touch. The signals these sensors transduce are mapped onto somatosensory maps or homunculi in the somatosensory receiving area (S1) and the secondary somatosensory cortex (S2), linking them to specific regions of the body (Dijkerman & De Haan, 2007). We do not directly experience four different kinds of pressure signals emanating from various regions of our skin. Instead, we perceive a tactile intentional object (Merleau-Ponty, 1945/2012), like a smooth table or a rough rock. Not only do these sensory stimuli meld together to form multi-modal perceptions, but they can alter our cross-modal perceptions. The McGurk effect is one well-known case of this, wherein vision (e.g., perceiving lip movements) can alter auditory perceptions (McGurk & MacDonald, 1976).

If a simulated environment fails to support these sensors (or the physical phenomena they are intended to receive), an agent's embodied experienced environment—its *Umwelt*—will necessarily be different. Yet, even ticks have a richer *Umwelt* than is supported by most of today's virtual environments. A tick's *Umwelt* results primarily from a combination of tactile hairs and Haller's organ. Hairs provide it with a basic sense of touch, allowing it to negotiate plants or the rough environments of hairy mammalian skin. And Haller's organ allows it to transduce information from airborne particles (olfaction), temperature, humidity, and light. Uexküll (2010) described the tick's *Umwelt* thus:

> The tick hangs inert on the tip of a branch in a forest clearing. Its position allows it to fall onto a mammal running past. From its entire environment, no stimulus penetrates the tick. But here comes a mammal, which the tick needs for the production of offspring. And now something miraculous happens. Of all the effects emanating from the mammal's body, only three become stimuli…From the enormous world surrounding the tick, three stimuli glow like signal lights in the darkness and serve as directional signs that lead the tick surely to its target. (Uexküll, 2010, p. 51)

The relative desolation of a tick's *Umwelt* starkly contrasts with the rich and colorful world of experience available to humans, and capturing this cornucopia of sensations in our simulated environments is a formidable challenge. Importantly, our sensors and *Umwelt* did not evolve for our spectatorial enjoyment. They evolved because they enhance our adaptive fit to our environment, and they serve the pragmatic function of guiding embodied action in the world.

## 2.3 Correspondence of Evolutionary and Experiential Histories

Natural agents are both phylogenetically and ontogenetically attuned to their environments (Gallagher, 2017). This involves a coevolution with their ecological niches (Odling-Smee et al., 2003; Varela et al., 1991/2016). This coevolution means that organisms do not merely fortuitously find an ecological niche and unilaterally adapt to it (Laland, 2017; Odling-Smee et al., 2003; Sterelny, 2003). They often take an active role in shaping their niches, making them more fitting to their needs. Beavers, for example, alter their niche by building dams. Ants construct elaborate mounds. And, more than any other animal, humans have in the last 12,000 years (counting since the Agricultural or Neolithic Revolution) radically reshaped their ecological niches, creating roads, farms, and cities.

The evolutionary heritage of any natural agent is a complex history of sedimented adaptations to changing environmental pressures. Human capacities, such as trichromatic vision, have roots in early primate adaptations to arboreal conditions, where trichromacy conferred an advantage in perceiving vibrant fruits against the background of green leaves (Osorio & Vorobyev, 1996). Many adaptive and maladaptive behaviors (as well as seemingly neutral proclivities) may have their roots in evolutionary processes. A prime example of this is the evolution of long-term sexual strategies and motivations (see Brase, 2006; Buss, 1994; Buss & Schmitt, 1993; Kenrick et al., 1996; Salska et al., 2008; Schulte-Hostedde et al., 2008; Schwarz & Hassebrauck, 2012; Shackelford et al., 2005; Smuts, 1995; Wade et al., 2009).

Modeling these ancestral attunements in software agents can be tricky. An agent's innate drives and motivations (such as survival, curiosity, and reproduction) provide the impetus for action, yet we do not *directly* know what those primal imperatives are. Similarly, most (if not all) natural agents are equipped by evolution with innate reflexes and other fixed action patterns. The rooting, sucking, and stepping reflexes exhibited by human babies are some examples. These evolutionary factors must be accounted for in our cognitive theories, as they inform what must be *built-in* (rather than learned) to support the development of human-like artificial minds. When studying humans, the circumstances of modern life, particularly social and cultural contexts, can make unearthing these hidden factors exceedingly difficult. Ancient evolutionary heritage can manifest in unexpected ways when environmental pressures and conditions change; for example, the modern prevalence of depression, anxiety, and hypertension may be intimately related to the new pressures that modern life places on humans.

These evolutionary attunements are shaped, refined, and added to through a lifetime of experiences. In other words, phylogenetic attunement or adaptability is complemented by ontogenetic attunement. The result is that each individual has a unique experiential trajectory that influences its behaviors. These experiences can manifest in the acquisition of beliefs, social and cultural norms, skills, languages, and a web of potentially complex motivations. Moreover, traumatic experiences (e.g., the death of a loved one, or physical and mental abuse) can irreparably and dramatically change the way agents perceive and interact with the world. Therefore, a fundamental challenge in modeling virtual humans is accounting for the behavioral influences exerted by these myriad experiences.

Beliefs, which typically develop from experience, exert a powerful influence on agential behavior. It has also been suggested that beliefs can modulate how entities, objects, and situations are perceived (Siegel, 2012, 2016; Stokes, 2013). Racial beliefs and attitudes, for example, can affect how humans perceive skin color (Levin & Banaji, 2006).[3] A complicating factor is that people often hold contradictory beliefs, or verbally advocate for one behavior while engaging in another. For example, many people consider themselves "pro-life" but also believe in the death penalty. While these two beliefs are not formally contradictory, they may lead to apparently inconsistent behaviors, and generate cognitive dissonance. Thus, the task is not only to model agents capable of mathematically



[3] While there is much empirical evidence for cognitive penetrability, the phenomenon has been subject to recent debates. See for example (Firestone & Scholl, 2016).

"optimal" or "rational" behaviors, but those capable of inconsistent, contradictory, and erratic behaviors that may seem at odds with their own interests and well-being.

These beliefs can be propositional or non-propositional. For example, human beings hold a myriad of non-propositional beliefs that can be made propositional *if needed*, but mostly operate as *dispositions to act* in an embodied and action-oriented fashion (Hornsby, 2012; Ryle, 1976). Knowledge regarding subtle cultural norms such as proxemics (i.e., how close it is acceptable to stand next to another person in various contexts) are largely a matter of non-propositional, low-level attunements to an environment. Importantly, many such attunements are not hardwired but develop through an agent's experiential history. Consequently, modeling human behavior requires that our virtual humans be capable of learning, adapting, and changing their beliefs and dispositions to act to stay sufficiently attuned to their environments.

While accounting for propositional beliefs in software agents may seem more tangible and manageable than non-propositional beliefs, attempts to reduce all human conduct to propositional form have met with limited success, and the engineering of such declarative knowledge can be monumentally time consuming[4]. Furthermore, language and the creation of software agents capable of thinking *in a language* is a formidable challenge given the richness and complexity that comes with language acquisition and use. Nevertheless, in order to faithfully model human behavior, we must overcome these technical challenges.

An important and common human activity that profoundly influences behavior and relies on linguistic thought is the generation of self-narratives. That is, human beings understand themselves and their place in the world through the lens of a self-generated story (Bruner, 2004; Dennett, 1992; Gallagher, 2020; Hutto, 2008; Schechtman, 1996). The narrative self is often developed along the lines of culturally, ethnically, and nationally defined genres (McAdams, 2006), and reflects subculture (Dickson & Wright, 2017), sexual orientation, gender (Compton, 2020; McLean et al., 2020; Nelson & Fivush, 2020), and numerous other categories. Self-narratives tend to incorporate the narratives of others, and some have even suggested that there is a constant recursive relationship between an agent's embodiment, their social interactions, their available affordances, and their self-narratives (Dings, 2019). Furthermore, humans often act in accordance with distal intentions, which largely develop through experience in the context of self-narratives. This requires that one not only model and implement the mechanisms for constructing coherent self-narratives but also the selection of action in accordance with those narratives. Without the capacity to generate and act in accordance with self-narratives, the long-term behaviors of our virtual humans will almost certainly diverge from that of real humans. Such narratives may also be useful for understanding the basic intentionality of other agents (Hutto, 2008).

In summary, human behavior depends on personal and evolutionary histories that are difficult to model in software. Evolutionary factors inform what must be built-in rather than learned to support the development of human-like artificial minds; however, it can be difficult to discern their existence and contributions to human behavior experimentally. These evolutionary forces are modified and augmented by a lifetime of personal events that can result in the acquisition of beliefs, social and



[4] The Cyc project, started in 1984, is a long-running attempt at hand-engineering "common sense" in software to facilitate the construction of *expert systems*. As of this writing, Cyc's knowledge base is said to contain "10,000 predicates, millions of collections and concepts, and more than 25 million assertions" (*Cyc's Knowledge Base – Cycorp Inc.*, n.d.). According to Cycorp, it has taken over *4 million hours* to develop this knowledge store and its associated inference engine.

cultural norms, skills, languages, and a web of potentially complex motivations. Accounting for these experiential forces in software agents is challenging because they require time-consuming developmental processes that are difficult to replicate *in silico*. Finally, propositional beliefs can further manifest in linguistic thought, including self-narratives. These self-narratives may serve numerous purposes, including the setting of distal intentions. The behaviors of virtual humans that lack the ability to generate and act in accordance with such distal intentions will likely diverge from those of real humans.

### 2.4 Correspondence of Social and Cultural Contexts

Perhaps the most challenging aspect of the real world to simulate in the virtual human methodology is the incorporation of realistic social and cultural contexts. For example, one of the most basic effects in this domain is the influence of social group size on individual behavior. A classic example of this is in the infamous murder of Kitty Genovese in New York in 1964. Although many reportedly heard her cries for help in this populous New York City neighborhood, not a single person intervened or called the police. Naive explanations tend to attribute this lack of intervention to callousness or selfishness, and indeed that is how the media reported it at the time (Ross & Nisbett, 2011). Yet a series of experiments by Latané and Darley soon revealed that the explanation is rather to be found in the effects of social groups themselves. There tends to be a diffusion of responsibility in large groups (Darley & Latané, 1968; Latané & Darley, 1969). Paradoxically, the increased presence of people around Kitty Genovese led to a lack of anyone intervening. Presumably, no one called the police because everyone thought someone else was surely calling.

Social group size is only one simple example of these phenomena. Human social systems are intertwined with complex cultural systems and institutions that pervasively modulate cognition and the brain. For example, if we want to accurately predict how an individual human might respond to a life stressor, we must know their culture. People in highly collectivist cultures frequently seek out social support in family and friends, while those in highly individualist cultures tend towards rumination and isolation. These culture-specific behavioral tendencies may help explain higher incidences of depression in Western cultures (Ross & Nisbett, 2011), which tend to be more individualistic. Individualist and collectivist differences may also help explain differences in perception. Something as basic as visual fixation patterns in scene perception can be affected by culture, with persons in individualist cultures tending to fixate more on the salient focus of a scene. In contrast, persons in collectivist cultures tend to fixate more on contextual features (Chua et al., 2005).

We must also be cognizant of the social, cultural, racial, and gender biases that researchers might inadvertently introduce into their models of human minds. For example, most psychological studies are conducted on WEIRD people (Henrich et al., 2010), that is, people from **W**estern, **E**ducated, **I**ndustrialized, **R**ich, and **D**emocratic societies. In contrast, the vast majority of *Homo sapiens* that have lived on this Earth over the past 300,000 years are decidedly *not* WEIRD. Moreover, that data is overwhelmingly from a specific subset of WEIRD culture: educated, undergraduate students. Since our best data about human behavior comes from such a highly skewed population, the behaviors of virtual humans constructed based on that data will likely be disproportionately biased towards the behaviors of WEIRD people.

### 2.5 Institutional Contexts

But the problem of culture runs far deeper than that. Institutions, practices, technologies, and people act as external scaffolds for many cognitive processes. These networks form cognitive institutions, "pieces of the mind, externalized in their specific time and place, and activated in ways that extend our cognitive processes when we engage with them" (Crisafi & Gallagher 2010, pp. 124–125). Consider a scientific cognitive institution like the Hubble Space Telescope (i.e., not just the physical satellite in orbit, but also the scientists and regulatory bodies involved). No single person discovers new information about the age of the universe. It is the Hubble cognitive institution as a whole that produces new discoveries (Giere, 2006). What an individual scientist knows and does is constrained by their colleagues, by political directives, social pressures, and the technology itself. For example, few non-scientists are aware that in order to use the Hubble Space Telescope research teams must put in lengthy applications that are subjected to intensely competitive review.[5] Similarly, unlike the distorted reality portrayed in most movies, the use of equipment, such as super computers, and the running of simulations, require that scientists apply months, or years, in advance. Knowledge and knowledge production are intimately tied into the bureaucracy and processes of human institutions.

It is not just the knowledge, rules, and procedures encoded in brains that determine cultural networks or cognitive institutions. There is a deep materiality to any cultural system. Culture is just as much material things—boxes, books, clothes, cars, houses, buildings, and food—as it is ideas. Both material culture and other aspects of the environment are nontrivially a part of cultural networks and cognitive institutions. Consider Oldowan and later traditions of prehistoric stone tools. The forms of these tools, and the behaviors needed to make them, reflect the shape and material properties of the stones from which they were crafted. Different stones afford different manufacturing opportunities, and their shapes constrain how they must be flaked or otherwise processed. Once manufactured, these tools bestowed upon their ancient makers more complex and efficient forms of hunting and food preparation, as well as the ability to manufacture other items of material culture, such as clothing. They also provided defense against aggressors, and likewise facilitated aggressive acts against others. In other words, material culture has the capacity to dramatically transform individual behaviors, social interactions, and can birth cognitive institutions. Such cultural artifacts shape the brains of those engaged in their making and use (see Malafouris, 2013) as surely as the toolmakers themselves shape materials from the environment into useful tools. Critically, these practices are not reducible to neural representations in the brain of any agent, but involve a dynamical coupling between agents, their material culture, and other aspects of their environment.

Contemporary social practices and interactions are heavily dependent upon the built environment and the material culture in which they are immersed. In courts of law, judges are often placed on a central and raised platform, directing audience attention and respect, and shaping the arrangement of legal proceedings. As many educators know, classroom dynamics can be transformed with a simple shift in desk arrangement. Social interaction is not merely a product of individual human brains or minds; rather it plays out in cultural environments where the material setup may be just as important as the more cognitive and neural factors involved.

The effects of informational isolation on cognitive institutions can also be extensive. Newcomb's (1943) studies of the geographical spread of ideas and practices in the context of a small liberal arts college in Vermont, Bennington College, provides a classic example. While a majority of students came from wealthy, conservative backgrounds, most of them quickly developed a strong liberal identity that persisted many decades after their collegiate experience (Alwin et al., 1991). The Bennington atmosphere was strong enough to overcome the tendency that political ideology has to propagate among familial lines. The prime factor in the political sway of the college was geographic



[5] https://www.nasa.gov/mission_pages/hubble/servicing/series/How_science_is_done.html

isolation: Bennington was relatively isolated from the larger communities from which these students originated. A general tendency towards liberal politics became amplified in the cloistered environment of the college, where social connections were predominantly between college members rather than with members of the broader community. The message of the study is not about politics, per se, but about the way geography constrains social networks. Today, many of the same cloistered, amplifying effects originate not from geographic isolation but from informational isolation. This is due to the ubiquity of internet algorithms and social networks. Twitter and Facebook bubbles have become so polarized, particularly in the United States, that entirely different narratives of the same events are propagated to different communities. Cognitive institutions always have borders, and those borders can, at times, be sharply defined by factors such as geography, online social network connections, and other factors.

Cognitive institutions are themselves connected to other cognitive institutions. For example, “we know that research questions and decisions in science are not determined purely by scientific procedure, and scientific results are not strictly confined to scientific labs” (Slaby & Gallagher, 2014, p. 5). The kinds of decisions that individual scientists make in a laboratory may be determined and constrained by political institutions, funding organizations, career expectations (e.g., the pressure to gain tenure), and financial market pressures.

Once social, cultural (including material culture), and institutional factors are considered, the prospects of accurately modeling individual human cognition become exponentially more difficult and complicated. Yet it is a simple fact that humans do not act in social and cultural voids. As John Donne’s great poem goes, “No man is an island, / Entire of itself, / Every man is a piece of the continent, / A part of the main.” But that continent is not just other people—it is built from the vast and multitudinous cognitive institutions that shape our lives, our behavior, and our minds, from the home (which is a deeply cultural institution), to school, to work, and play. Diverse cultural practices and norms transform the way individuals approach and understand the world. Specific cognitive institutions shape cognition and behavior. Simulating humans without simulating social, cultural, and institutional contexts will result in a one-sided and skewed model of real human cognition and behavior. Virtual humans will be like “islands” divorced from the continent of which they are a part without these contexts.

Despite the existence of social simulations, they remain simulations of population-level phenomena. The computational capabilities needed for institutional or societal simulations are far beyond any current technology. Even while current social simulations prove promising models of macrosocial patterns, they do not model individual, *personal*, human agents interacting in a social milieu. Existing models include those that are essentially the progeny of the classical Lotka-Volterra equations modeling predator-prey populations (Abdollahian et al., 2013). They are systems of coupled dynamical equations that model macroscopic trends, not individual people. Sociological simulations have been pioneered by Bainbridge (1987, 1995), who uses neural nets to simulate religious belief in multi-agent systems. While this approach captures much more relating to particular agents’ individuality, it is still nothing like a full simulation of virtual humans in a virtual environment. And although archaeologists and anthropologists have begun using agent-based and systems-dynamics models to model everything from Neolithic cultural patterns (Shults & Wildman, 2018) to the transmission of early Christian rituals (Kaše et al., 2018), as useful as these simulations may be, none of them come close to an immersive virtual human simulation.

## 3 Alternative Methodologies Not Based on Correspondence

In Section 2, we presented a critical examination of the virtual human research methodology. Specifically, we described several correspondence problems that can arise when attempting to replicate human minds (or other complex natural minds) *in silico*. As a result of these problems, we believe the goals of the virtual human methodology are currently unrealistic. Moreover, the path for achieving those goals is ill-defined. These challenges suggest the need for viable alternative synthetic methodologies with more realistic goals. Specifically, we advocate for approaches that (1) are compatible with our current capabilities, (2) cultivate the *incremental* acquisition of enabling technologies and engineering competences, and (3) enrich our foundational understanding of minds and environments. We consider two such approaches here.

The first of these is the *classical animat approach* (see Section 3.1), which begins by constructing simple virtual ecological niches and autonomous agents with biologically inspired needs (e.g., survival and reproduction). These autonomous agents are referred to as *animats*. As animats that can survive and thrive within these virtual ecological niches are discovered, they are subjected to more demanding environmental conditions. Animat complexity is *gradually* increased until new "species" of animat that can cope with these emerging environmental challenges are discovered. This process continues *ad infinitum*. Critically, animats are evaluated based on their ability to satisfy their own needs, *not on their resemblance to a natural species*. Therefore, the classical animat approach avoids the correspondence problems that frustrate the virtual human methodology.

The second is a *theory-driven animat approach* (see Section 3.2), which augments the classical animat approach with a design heuristic based on a cognitive architecture (see Section 3.2.1). The utility of this approach is that it allows a cognitive theory to dictate the permissible animat designs without resorting to a resemblance-based evaluation criterion (i.e., one that is based on the perceived degree of similarity between natural and engineered systems). This may, for example, increase the likelihood that animats will be discovered with more human-like minds. However, the resulting biases may also prevent the discovery of promising mechanisms of mind based on different principles. Therefore, this approach may not always be preferable to the classical animat approach in practice. Like the classical animat approach, the theory-driven animat approach begins with simple agents and environments, and gradually increases their complexity. We illustrate this methodology using the LIDA (Learning Intelligent Decision Agent) cognitive architecture in Section 3.2.2.

The choice of whether to apply the classical or theory-driven animat approach is analogous to the choice between divergent and convergent modes of ideation (see Cropley, 2006). If one wants unconstrained access to all possible mechanisms of mind, then the classical approach is to be preferred. This may be particularly useful when surveying or comparing a set of animat designs *in search of a theory*. As a trade-off, there are no selective pressures built into this methodology that promote the creation of human-like minds; it delivers intelligences with increasing capabilities. In contrast, the theory-driven animat approach sacrifices the full breadth of animat possibilities in the hope of expediting the discovery of more sophisticated, naturalistic and human-like animats. However, the theoretical biases and constraints introduced using this approach may ultimately impede progress and discovery if one's assumptions are unsound. In practice, it may be beneficial to switch back and forth between the two approaches (becoming less or more constrained) as circumstances dictate.

Both of these approaches are viable in practice because they begin with *simple* environments and autonomous agents, and they do not require that the resulting animats resemble a natural species. A basic feature of these approaches is that the initial environments and agents are constrained to be possible under current engineering practice. Assuming subsequent iterations of the animat approach are based on "minimal" increases in environmental and agential complexity, then, in theory, the approach should converge on animats that closely reflect the limits of one's own technological and engineering capabilities. Once those limitations are identified, it may be possible to address them in a deliberate and targeted fashion.

## 3.1 The Classical Animat Approach

The animat methodology (see S. W. Wilson, 1991) proposes that one begins the investigation of minds by creating *simple* autonomous agents (i.e., animats) that are embedded in naturalistic environments. These autonomous agents are primarily focused on satisfying somatic (e.g., obtaining sustenance), homeostatic (e.g., maintaining comfortable body temperatures), and reproductive drives (e.g., mating and rearing young), as well as other, more derivative, survival-oriented needs (e.g., curiosity and social acceptance). Stewart Wilson (1991) argued that modeling these intrinsic motivations is essential since they are likely the "principal drivers" of behavior, and shape how agents perceive, and conceive of, their worlds.

Once an animat of minimal complexity is created, one gradually increases its complexity in response to more demanding environments and more exacting survival-oriented needs:

> given an environment and an animat with needs and a sensory/motor system that satisfies these needs to some criterion, increase the difficulty of the environment or the complexity of the needs—and find the minimum increase in animat complexity necessary to satisfy the needs to the same criterion. (S. W. Wilson, 1991, p. 16)

After many such iterations, the goal of this methodology is for these animats to become more capable and sophisticated artificial minds, and their environments more complex and challenging.

An animat's "quality" is judged by its ability to enact behaviors that allow it to survive and thrive within a virtual ecological niche. While the use of naturalistic environments and survival-oriented drives *may* foster the development of animats that resemble some natural species, the animat approach is itself agnostic to the constitution of these artificial minds. As a result, it avoids the correspondence problems that complicate the virtual human methodology.

As a trade-off, there are no selective pressures built into this methodology that promote the creation of human-like minds. Therefore, even if the approach converges on sophisticated artificial minds capable of "general" intelligence, *it offers no guarantees that the resulting minds will be human-like*. Consequently, the goals of the classical animat approach are different from those of the virtual human methodology. This shift in objective is necessary because the goals of the virtual human methodology are currently unrealistic. Nevertheless, the knowledge and engineering capabilities discovered while applying the classical animat approach may enable the virtual human methodology in the future. (We return to this idea in Section 4.)

The use of virtual ecological niches and naturalistic drives differentiate the classical animat approach from the performance-based approaches that dominate mainstream artificial intelligence research. As such, the animat approach falls within the domain of Alife simulations (e.g., Varela, 1988).

### 3.2 A Theory-Driven Animat Approach

The animat approach described in Section 3.1 avoids the correspondence issues described in Section 2 by permitting *any* mechanism of mind that generates adaptive behaviors within some virtual ecological niche. The consequence of this unconstrained exploration of artificial minds is that the resulting minds may not be human-like or even animal-like. For some, this may not be an issue. Langton (1997) mused, "Artificial Life need not merely attempt to recreate nature as it is, but is free to explore nature as it could have been" (Langton, 1997, p. x). Nevertheless, for cognitive scientists that are primarily interested in human intelligence, the compromises required by the classical animat approach may be unacceptable.

In this section, we speculate on the possibility of using a cognitive architecture as a *heuristic* to guide the incremental selection of animats towards those with more human-like intelligence. As with all heuristics, it is not guaranteed to work in practice, and its value is only as good as the validity of one's assumptions about the nature and composition of human minds. The utility of this approach is that it allows a cognitive theory to dictate the permissible animat designs *without resorting to a resemblance-based evaluation criterion*, which is impossible to apply in current practice. This approach merely constrains the permissible animats to the region of animat design space consistent with the chosen cognitive architecture.

This *theory-driven animat approach*, like the classical animat approach, is bottom-up and incremental. It starts with simple environments and agents, and gradually scales up their complexity. Furthermore, like the classical approach, it does not depend on validating that the resulting animats have human-like minds. Animats are judged solely on their ability to satisfy their own needs within a virtual ecological niche. We elaborate further on this approach in the subsections that follow.

#### 3.2.1 Unified Theories of Cognition and LIDA

Many systems-level cognitive architectures (see Kotseruba & Tsotsos, 2018) strive to be "unified theories of cognition" (Newell, 1994) that are capable of modeling many, if not all, human cognitive activities and processes. Cognition, in this sense, broadly encompasses every mechanism of mind, including (but not limited to) perception, motivations, action selection, motor control, attention, learning, metacognition, sense of body and self, and language. Biologically inspired cognitive architectures (BICAs), such as LIDA (Learning Intelligent Decision Agent; see Franklin et al., 2016), additionally constrain artificially intelligent systems to be more like their natural counterparts, based on our current beliefs about natural minds.

Among the BICAs, LIDA is particularly well-suited to serve as a theory for guiding the creation of incrementally more human-like animats, for the following reasons:

(1) LIDA has a well-developed motivational system (McCall et al., 2020) that supports and modulates its many cognitive processes, including action selection and learning. This accords with the animat approach's emphasis on survival-oriented needs being the primary drivers of behavior.
(2) LIDA has a highly modular design with a multitude of distinct short- and long-term memory modules, and supporting cognitive processes (see Figure 2). This modularity turns out to be very useful for designing animats of varying capabilities and complexity (see Section 3.2.2).
(3) LIDA implements and fleshes out many psychological theories (Baddeley & Hitch, 1974; Barsalou, 1999; Conway, 2001; Ericsson & Kintsch, 1995) including the Global Workspace Theory (Baars, 1988) of consciousness. While the scientific study of consciousness has

become more acceptable in recent years, research in machine consciousness and the attempted construction of conscious artifacts (Franklin, 2003) has been largely neglected. Accounting for consciousness is an important aspect of modeling human-like minds, and it has been rarely attempted in a cognitive architecture.

(4) LIDA is an embodied cognitive architecture, incorporating situated cognition and grounded representations, and adhering to the principle that cognition is primarily for action (Franklin, 1995, Chapter 16; M. Wilson, 2002). These features collectively endow LIDA agents with the potential of operating within a wider range of complex naturalistic environments than cognitive architectures that are more specialized towards symbolic environments and tasks.

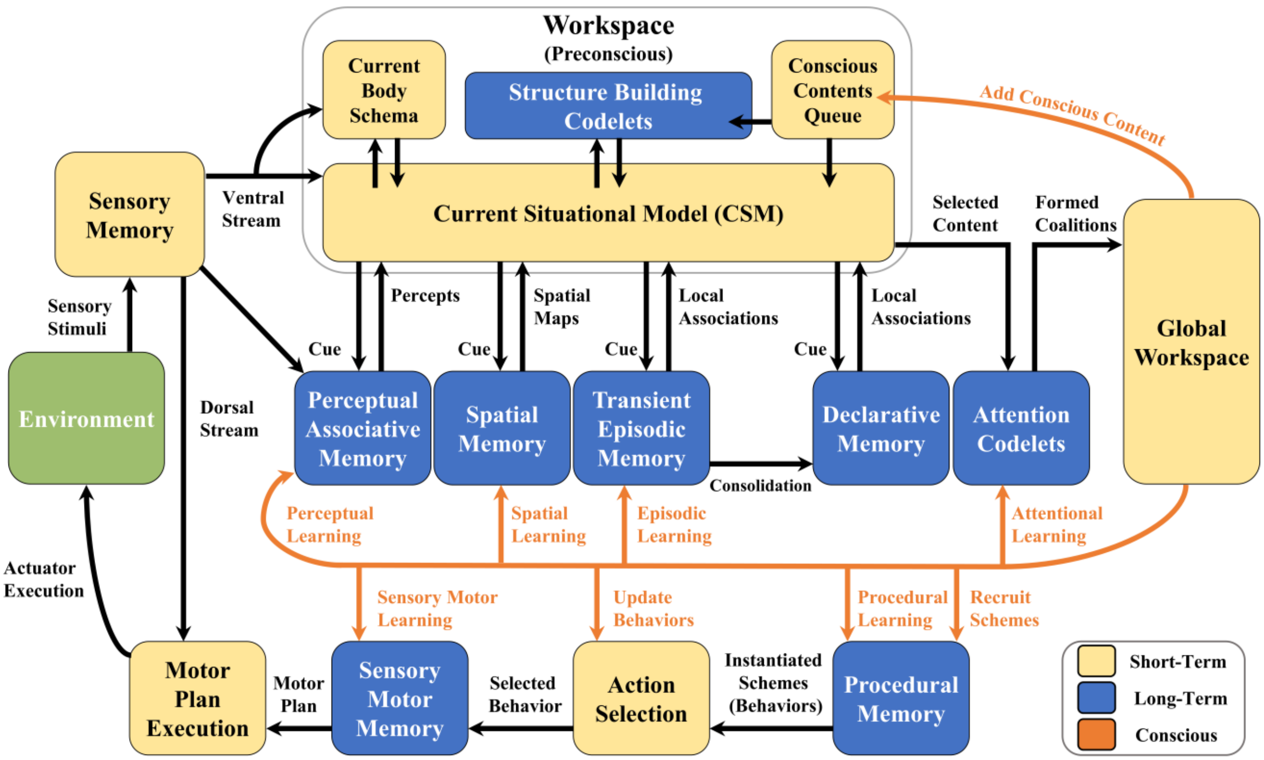


**Figure 2. The LIDA cognitive cycle.**

Learning Intelligent Decision Agent (LIDA) is composed of many short- and long-term memory modules, codelets (i.e., special-purpose processors), and supporting cognitive processes (e.g., consolidation, cueing, learning, and decay). All cognitive activities and processes are conceptualized as occurring within, or emerging as the result of, a continual series of potentially overlapping *cognitive cycles*[6]. Cognitive cycles are viewed as being sub-divided into three phases: (1) perception and understanding, (2) attention, and (3) action and learning.

During LIDA's *perception and understanding phase*, sensory stimuli from an agent's environment can activate low-level feature detectors in Sensory Memory. These, in turn, can activate perceptual



[6] The cognitive cycle corresponds to the "action-perception cycle" referred to by many psychologists and neuroscientists (see Freeman, 2002; Fuster, 2004; Neisser, 1976).

representations[7] in Perceptual Associative Memory (PAM). Perceptual representations receiving sufficient activation (from Sensory Memory and other representations in PAM) are instantiated[8] as *percepts* in the Current Situational Model (CSM)—a sub-module of LIDA's Workspace. These percepts may correspond to recognized objects, entities, situations, and events, as well as their associated affective content (e.g., feelings, emotions, desires, and dreads; see McCall et al., 2020). In addition to percepts, the CSM also receives sensory content from Sensory Memory and the Current Body Schema (see Neemeh et al., 2021). Structure building codelets operate on the representations in the CSM, creating new associations (e.g., causality links) as well as more complex structures. These can include event structures, mental simulations (see Kugele & Franklin, 2020), spatial maps (see Madl et al., 2018), self-narratives and distal intentions (see Kronsted et al., forthcoming), and plans, among other things. The representations in the CSM may also *cue* associated long-term memories (e.g., episodes and semantic memories) into the CSM. The representations contained within the CSM correspond to an agent's *preconscious*[9] understanding, interpretation, and "thoughts" pertaining to its current situation.

During LIDA's *attention phase*, attention codelets can identify preconscious representations in the CSM that are of interest to them based on their own concerns (e.g., brightness, loudness, novelty, surprise, or urgency). If such content is found, an attention codelet will bring it to a "coalition forming process," which may create a *coalition* that includes that codelet and the content it promotes. Coalitions compete in a winner-take-all competition in the Global Workspace based solely on the coalitions' activations. The winning coalition and its content are globally broadcast to all of LIDA's modules. The content in the global broadcast is said to be "functionally conscious."[10]

During LIDA's *action and learning phase*, content from the global (conscious) broadcast is received by all modules, including Procedural Memory, which uses that content to activate and instantiate its *schemes*. Schemes are representations that correspond to consciously observed correlations between (situational) contexts, actions, and the results of those actions in those contexts. Each scheme additionally has a *base-level activation* that estimates the likelihood that the agent's actions will produce the scheme's expected results when executed in similar contexts. Instantiated schemes are referred to as *behaviors*. Behaviors receiving sufficient activation are sent to LIDA's Action Selection module to compete as candidates for an agent's next selected behavior. Action Selection chooses (at most) one of its behaviors per cognitive cycle (which may include non-decayed behaviors from a previous cognitive cycle) to be its currently *selected behavior*. It then sends this selected behavior to LIDA's Sensory Motor System (SMS; Dong & Franklin, 2015) for execution. The SMS is composed of two modules: Sensory Motor Memory (SMM) and Motor Plan Execution (MPE).



[7] LIDA is a hybrid cognitive architecture that can be described as including both symbolic and non-symbolic representations, as well as non-representational modules (e.g., its Sensory Motor System). The existence and nature of mental representations in natural systems (e.g., brains) remains a contentious and highly debated topic in cognitive science, and some of the authors of this article argue against them (see, e.g., Gallagher, 2017).

[8] Instantiation is the process by which specific concrete instances are generated from more general templates by binding values to unspecified variables and parameters. For example, schemes in Procedural Memory are instantiated into behaviors by binding free variables in a scheme's context, action, or results. Where an uninstantiated scheme may contain a generic `OBJECT` placeholder variable, the instantiated scheme (i.e., behavior) would replace `OBJECT` by a specific object from the current global broadcast (e.g., a `CHAIR`). A similar process of instantiation occurs when Perceptual Associative Memory instantiates percepts, and Sensory Motor Memory instantiates motor plans.

[9] We use the convention established by Franklin and Baars (2010) of referring to unconscious representations that have the *potential* to become conscious as "preconscious" and those that do not as "never-conscious."

[10] LIDA currently makes no claims regarding phenomenal consciousness.

SMM is a long-term memory module that instantiates *motor plan templates* into *motor plans* based on a selected behavior. MPE executes motor plans through a process of situated, “online control,” during which, *motor commands* (i.e., low-level directives) are sent to an agent’s actuators in response to its immediate “situated” concerns.

LIDA’s numerous learning mechanisms (see Kugele & Franklin, 2021) can also be invoked during the action and learning phase, as a direct result of a conscious broadcast. These mechanisms support the learning of new representations, and the reinforcement of previously learned representations.

For a more comprehensive introduction to LIDA, see Franklin et al. (2016).

### 3.2.2 An Illustration of the Theory-Driven Animat Approach using LIDA

In this section, we illustrate how a unified theory of cognition, specifically LIDA, *might* inform design choices at each step when applying an animat-style research methodology. The advantage of doing so is that such a *theory-driven animat approach* has the potential to constrain these engineered autonomous agents to be more like natural systems (e.g., human and non-human animals) than their unconstrained counterparts. The described progression extends from *minimal reactive agents*, to *minimal conscious agents*, and beyond. While we focus here on illustrating the gradual addition of modules and processes, refinements *within* each module and process may be equally important in practice.

As we have previously stated, this approach, like the classical animat approach, is not dependent on validating that the resulting engineered species resemble any natural species. This is in sharp contrast with the virtual human methodology. Instead, the theory-driven animat approach is focused on expanding our foundational understanding of mechanisms of mind rather than replicating natural minds *in silico*.

A *minimal reactive agent* (see Figure 3, Panel 1) could be implemented using LIDA’s Sensory Memory and Motor Plan Execution modules. Such agents have a single motor plan that emits motor commands based solely on incoming sensory stimuli and Sensory Memory’s activated low-level feature detectors. While these agents are incapable of “offline” cognitive activities (e.g., reasoning, introspection, and the recall and formation of long-term memories; see M. Wilson, 2002), this form of purely situated control may be sufficient for extremely simple agents and ecological niches.

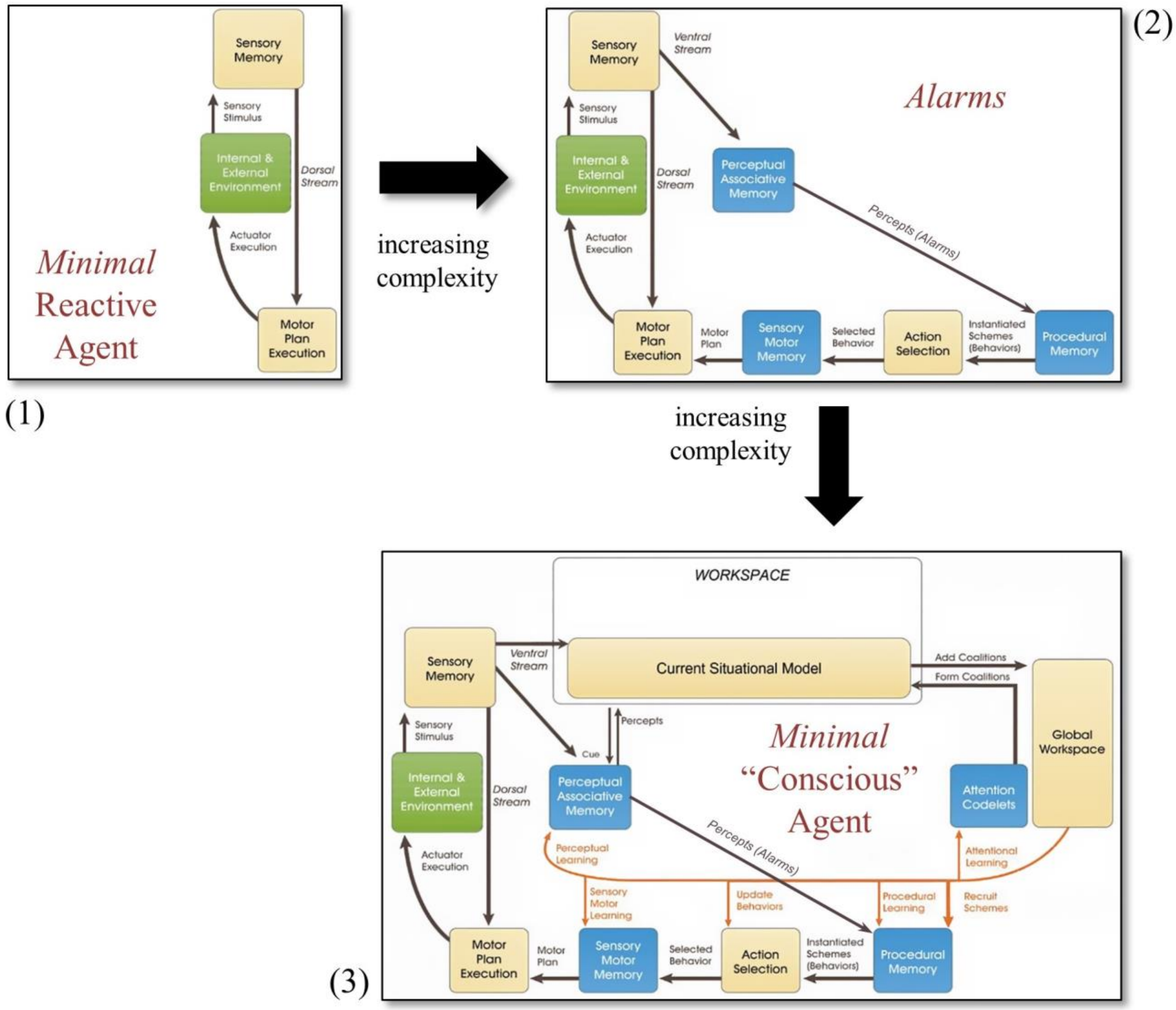


**Figure 3. An incremental progression of LIDA agent complexity. (1) shows a minimal agent control structure that operates solely through situated, online control (no offline cognition). (2) shows the addition of Perceptual Associative Memory and Procedural Memory modules and an Action Selection module operating via never-conscious "alarms." (3) shows a minimal "conscious" LIDA agent control structure that is capable of attention, conscious experiences, learning, and the consciously mediated selection of actions.**

Adding *alarms* (see Figure 3, Panel 2) greatly increases the flexibility of these simple reactive agents. Sloman (2001) referred to an alarm as a "purely reactive and pattern driven" (Sloman, 2001, p. 188) mechanism capable of simple behaviors such as freezing, fleeing, and aggressive displays. Alarms require the ability to *recognize* urgent situations and to *select* appropriate behavioral responses. These additional capabilities are supported by minimal implementations of Perceptual Associative Memory

(PAM), Procedural Memory, and Sensory Motor Memory (SMM). PAM instantiates alarm *percepts*[11] based on sensory stimuli that are recognized as demanding immediate reactions (e.g., life-threatening events and conditions). From these percepts, Procedural Memory instantiates an appropriate reactive behavior (e.g., fight, flight, or orienting response), and SMM instantiates a corresponding motor plan for situated execution. These agents are still unable to learn or engage in most offline cognitive activities (apart from simple long-term memory recall). In agents with more sophisticated cognitive capabilities (such as reasoning and deliberative action selection), alarms provide a “short-circuit” for bypassing these slower control mechanisms in situations that require very rapid reactions. For example, a driver may unconsciously engage the brakes and turn the steering wheel of their car in response to a vehicle suddenly swerving into the lane in front of them. The selection and execution of these emergency maneuvers often occurs prior to, or at the same time as, the conscious awareness of the alarm situation that inspired their selection.

A *minimal “conscious” agent* (see Figure 3, Panel 3) could be implemented by adding a Workspace (preconscious), Global Workspace, and one or more attention codelets. The introduction of conscious broadcasts sets the stage for a number of different learning mechanisms, including perceptual, procedural, sensory motor, and attentional learning. This class of agents also benefits from *consciously mediated action selection*, which allows the selection of actions, and the instantiation of motor plans, that are more attuned to the most salient aspects of their situational contexts. While their offline cognitive abilities are still quite limited, the introduction of associative and non-associative learning mechanisms, and consciously mediated action selection, would likely be highly adaptive in most environments.

At this point, there are many possible continuations depending on the needs of an agent and its environmental pressures. Adding a Current Body Schema would allow an agent to have a better sense of its current somatosensory inputs and improve the perception of action opportunities in its environment; Adding a Transient Episodic Memory would allow an agent to store and retrieve recent autobiographical episodes; Adding one or more structure building codelets would enable a wide variety of cognitive abilities, such as categorization, causality, planning, and mental simulation; And adding Spatial Memory would empower an agent with the ability to create cognitive maps (e.g., spatial maps) of portions of its environment. Each of these potential branching points result in many other choices that could be incrementally explored.



[11] An important class of these percepts are “feelings” (see McCall et al., 2020), which are affective appraisals that reflect an agent’s basic drives and motivators.

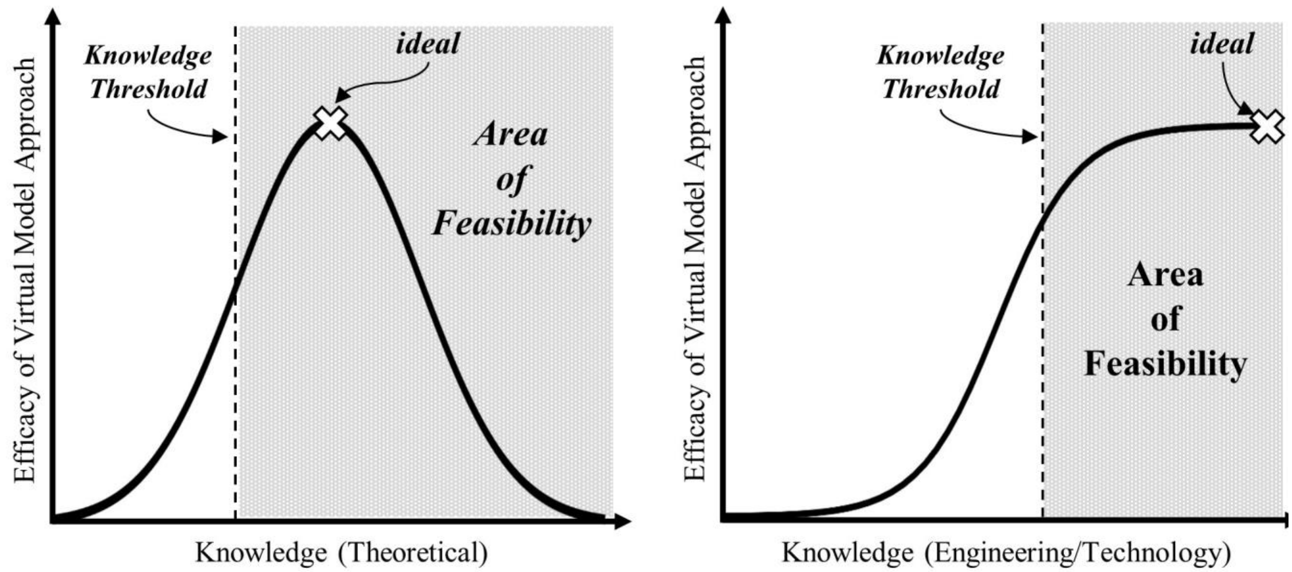


**Figure 3. The efficacy and feasibility of virtual model (e.g., virtual human) methodologies as functions of our theoretical (Left Panel) and engineering (Right Panel) knowledge. The shapes of these curves indicate different knowledge-efficacy relationships. Theoretical knowledge influences efficacy in a bell-shaped (or inverted-"U") relationship. Efficacy peaks when our theoretical knowledge is counter-balanced by remaining lines of scientific inquiry. If our theoretical knowledge about a target species (e.g., real humans) is limited, then our virtual models will lack external validity, rendering them useless. On the other hand, being extremely knowledgeable about a target species reduces the available lines of scientific inquiry, and once again limits the usefulness of our virtual models. Additional engineering knowledge (and supporting technologies) is always a facilitator; however, the impact of new engineering knowledge and technologies is greatly diminished once we are capable of creating sufficiently realistic bodies, environments, and mechanisms of mind.**

## 4 Discussion

The virtual human research methodology is a synthetic approach to understanding minds based on the creation of artificial, human-like minds that control virtual, human-like bodies in simulated worlds. The feasibility and efficacy of this, or any virtual approach that seeks to *replicate* natural minds *in silico*, depends on the right combination of theoretical and engineering knowledge (see Figure 4), and the availability of enabling technologies. A lack of theoretical knowledge leads to inaccurate models and ineffectual virtual minds. A lack of engineering knowledge leads to impoverished virtual environments, and an inability to manifest our models in software.

Unfortunately, at this moment in history, we lack in both an adequate theoretical understanding of human minds and the engineering know-how needed to create virtual humans and realistic simulated environments. The resulting disparities between simulation and reality will undermine the virtual human methodology and any attempts to generalize experimental results from virtual to real humans. Behavioral mimicry is not enough. Vision alone is not enough. And human-like minds (natural and artificial) cannot be considered separately from their environments; their experiences of those environments; their personal and evolutionary histories with those environments; and the social, cultural, and institutional contexts that occur within those environments.

One of the most problematic features of the virtual human methodology is not in its aspirations, but in its approach to achieving them. By focusing exclusively on the most complex of known organisms (i.e., humans), its progress is stymied from the start. In contrast, the alternative synthetic approaches presented in Section 3 implicitly assume that one must understand simpler minds and environments before embarking on the creation of complex minds and realistic virtual worlds. They also implicitly assume that what is learned from the creation of these simpler minds and environments will translate into knowledge that will facilitate the creation of more complex minds and environments. As such, these approaches provide a mechanism for *gradually* acquiring and refining the needed technologies and engineering competences. The virtual human methodology does not.

Another important feature of these alternative approaches is that they replace the *resemblance-based* evaluation criterion used in the virtual human research methodology (i.e., one that is based on the perceived degree of similarity between natural and engineered agents) with a *performance-based* criterion that judges autonomous agents based on their ability to produce adaptive behaviors. In other words, *these alternative approaches do not require that the engineered autonomous agents resemble any natural species*. This change in evaluation criterion is how these approaches avoid the correspondence problems introduced in Section 2.

While the goals of the virtual human methodology and these alternative synthetic methodologies are different, they are not orthogonal. Both seek to better understand minds and the mechanisms underpinning adaptive behavior. The virtual human methodology pursues these goals narrowly, focusing solely on explicating human intelligence through the creation of human-like autonomous agents. The animat-based approaches pursue these goals more broadly, admitting many, potentially disparate, mechanisms of adaptive behavior. Unlike the virtual human methodology, the primary goal of these alternative synthetic approaches is the expansion of our foundational understanding of minds and environments rather than replicating natural minds *in silico*. The generality of this goal is advantageous, as the resulting engineering and theoretical knowledge is likely to benefit all synthetic methodologies, including the virtual human methodology.

A natural question one might raise is: Do we have sufficient theoretical and engineering knowledge to replicate “simple” animals *in silico*, and experiment on them in lieu of their natural counterparts? In other words, is *any* virtual animal-based methodology feasible in current practice? Such an approach would still be subjected to many of the correspondence issues introduced in Section 2; however, one might hope that the bar would be sufficiently lowered to mitigate the most serious of these issues.

Let us consider *Caenorhabditis elegans*, a species of nematode (i.e., roundworm). We have a considerable amount of knowledge about the biology of *C. elegans*, largely due to the fact that it has been widely used as an animal model since the 1970s. Its entire genome has been sequenced, and it is the only organism to have a completely mapped connectome (i.e., neural wiring diagram). Compared with *H. sapiens*, *C. elegans* is extraordinarily simple. It has a few hundred neurons (compared to approximately 100 billion in *H. sapiens*) and a few thousand synaptic connections (compared to approximately 100-500 trillion in *H. sapiens*). In total, its entire body is composed of less than 1000 cells. Despite its neural simplicity, it has chemoreceptors, thermoreceptors, mechanoreceptors, nociceptors, and photoreceptors. It is capable of a multitude of behaviors. And it exhibits a surprisingly varied set of learning mechanisms (see Qin & Wheeler, 2007; Rankin, 2004).

Our extensive theoretical knowledge about *C. elegans*, combined with the relative simplicity of their bodies, environments, and *Umwelt*, make them a compelling starting place for developing our virtual animal modeling “chops.” And yet, simulating a *C. elegans* and its environment *in silico*, has proven to be an extremely difficult task. There have been numerous attempts (Blau et al., 2014; Gleeson et al., 2018; Kitano et al., 1998; Sarma et al., 2018; Suzuki et al., 2005; Szigeti et al., 2014) but no resounding successes, and none of these come close to modeling the full breadth of *C. elegans* behaviors or its environment. Even the basic neurobiology of its locomotion remains a mystery (Gjorgjieva et al., 2014). A common refrain in this literature is the fundamental difficulty of the task, and an appreciation for the limitations of our current knowledge. For example, Blau et al. (2014) stated,

> Caenorhabditis elegans features one of the simplest nervous systems in nature, yet its biological information processing still evades our complete understanding. The position of its 302 neurons and almost its entire connectome has been mapped. However, there is only sparse knowledge on how its nervous system codes for its rich behavioral repertoire. (Blau et al., 2014, p. 436)

The challenges inherent in any virtual animal approach, even one based on extremely simple species like *C. elegans*, is hard to overstate. Moreover, these birthing pains are only the beginning. The more daunting task may come when researchers attempt to substantiate claims that virtual and real *C. elegans* are similar enough to support scientific discovery. While we remain optimistic that such approaches are possible, they are extraordinarily difficult to realize in practice. Therefore, alternative synthetic approaches (such as those presented in Section 3), which are not based on replicating natural species *in silico*, may be necessary for the foreseeable future.

## 5 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## 6 References

Abdollahian, M., Yang, Z., Coan, T., & Yesilada, B. (2013). Human development dynamics: An agent based simulation of macro social systems and individual heterogeneous evolutionary games. *Complex Adaptive Systems Modeling*, *1*(1), 18. https://doi.org/10.1186/2194-3206-1-18

Afzal, A., Katz, D. S., Goues, C. L., & Timperley, C. S. (2020). A study on the challenges of using robotics simulators for testing. *arXiv*. https://arxiv.org/abs/2004.07368

Alwin, D. F., Cohen, R. L., & Newcomb, T. M. (1991). *Political attitudes over the life span: The Bennington women after fifty years*. Univ of Wisconsin Press.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, *25*(6), Article 6. https://doi.org/10.1038/s41591-019-0447-x

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Baddeley, A., & Hitch, G. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). Academic Press.

Bainbridge, W. S. (1987). *Sociology laboratory*. Wadsworth.

Bainbridge, W. S. (1995). Neural network models of religious belief. *Sociological Perspectives*, *38*(4), 483–495. https://doi.org/10.2307/1389269

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660.

Blau, A., Callaly, F., Cawley, S., Coffey, A., De Mauro, A., Epelde, G., Ferrara, L., Krewer, F., Liberale, C., & Machado, P. (2014). The si elegans project–the challenges and prospects of emulating caenorhabditis elegans. *Conference on Biomimetic and Biohybrid Systems*, 436–438.

Brase, G. L. (2006). Cues of parental investment as a factor in attractiveness. *Evolution and Human Behavior*, *27*(2), 145–157. https://doi.org/10.1016/j.evolhumbehav.2005.06.003

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *ArXiv*. http://arxiv.org/abs/2005.14165

Bruner, J. (2004). Life as narrative. *Social Research*, *71*(3), 691–710.

Buss, D. M. (1994). The strategies of human mating. *American Scientist*, *82*(3), 238–249.

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review*, *100*(2), 204. https://doi.org/10.1037/0033-295X.100.2.204

Campbell, M., Hoane Jr, A. J., & Hsu, F. (2002). Deep blue. *Artificial Intelligence*, *134*(1–2), 57–83.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *PNAS*, *102*(35), 12629–1233. https://doi.org/10.1073/pnas.0506162102

Compton, C. (2020). Co-sexuality and organizing: The master narrative of “normal” sexuality in the Midwestern workplace. *Journal of Homosexuality*, *67*(7), 1013–1039. https://doi.org/10.1080/00918369.2019.1582220

Conway, M. A. (2001). Sensory–perceptual episodic memory and its context: Autobiographical memory. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *356*(1413), 1375–1384.

Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal*, *18*(3), 391–404.

*Cyc’s knowledge base – Cycorp inc.* (n.d.). Retrieved December 15, 2020, from https://www.cyc.com/archives/service/cyc-knowledge-base

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4p1), 377.

Dennett, D. (1992). The self as a center of narrative gravity. *Self and Consciousness: Multiple Perspectives*, 111–123.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://arxiv.org/abs/1810.04805

Dickson, T., & Wright, R. (2017). The funny side of drug dealing: Risk, humor, and narrative identity. *Criminology*, *55*(3), 691–720. https://doi.org/10.1111/1745-9125.12148

Dijkerman, H. C., & De Haan, E. H. (2007). Somatosensory processing subserving perception and action: Dissociations, interactions, and integration. *Behavioral and Brain Sciences*, *30*(2), 224–230.

Dings, R. (2019). The dynamic and recursive interplay of embodiment and narrative identity. *Philosophical Psychology*, *32*(2), 186–210. https://doi.org/10.1080/09515089.2018.1548698

DiPaola, S. R., Bernardet, U., & Gratch, J. (Eds.). (2021). Modeling Virtual Humans for Understanding the Mind [Special Issue]. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/research-topics/13415/modeling-virtual-humans-for-understanding-the-mind/overview

Dong, D., & Franklin, S. (2015). A new action execution module for the learning intelligent distribution agent (LIDA): The sensory motor system. *Cognitive Computation*, *7*(5), 552–568.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., . Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, *31*, 59–79.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for “top-down” effects. *Behavioral and Brain Sciences*, *39*. https://doi.org/10.1017/S0140525X15000965

Franklin, S. (1995). *Artificial minds*. MIT Press.

Franklin, S. (2003). IDA, a conscious artifact? *Journal of Consciousness Studies*, *10*(4–5), 47–66.

Franklin, S., & Baars, B. (2010). Two varieties of unconscious processes. In E. Perry, D. Collerton, H. Ashton, & F. LeBeau (Eds.), *New horizons in the neuroscience of consciousness* (pp. 91–102). John Benjamins.

Franklin, S., & Graesser, A. (1997). Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages* (pp. 21–35). Springer-Verlag.

Franklin, S., Madl, T., Strain, S., Faghihi, U., Dong, D., Kugele, S., Snaider, J., Agrawal, P., & Chen, S. (2016). A LIDA cognitive model tutorial. *Biologically Inspired Cognitive Architectures*, *16*, 105–130.

Freeman, W. J. (2002). The limbic action-perception cycle controlling goal-directed animal behavior. *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN’02 (Cat. No. 02CH37290)*, *3*, 2249–2254.

Fuster, J. M. (2004). Upper processing stages of the perception–action cycle. *Trends in Cognitive Sciences*, *8*, 143–145.

Gallagher, S. (2013). The socially extended mind. *Cognitive Systems Research*, *25*, 4–12. https://doi.org/10.1016/j.cogsys.2013.03.008

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind*. Oxford University Press.

Gallagher, S. (2020). *Action and interaction*. Oxford University Press.

Gallagher, S., Mastrogiorgio, A., & Petracca, E. (2019). Economic reasoning and interaction in socially extended market institutions. *Frontiers in Psychology*, *10*, 1856. https://doi.org/10.3389/fpsyg.2019.01856

Geerts, H. (2009). Of mice and men. *CNS Drugs*, *23*(11), 915–926.

Giere, R. N. (2006). The role of agency in distributed cognitive systems. *Philosophy of Science*, *73*(5), 710–719. https://doi.org/10.1086/518772

Gjorgjieva, J., Biron, D., & Haspel, G. (2014). Neurobiology of Caenorhabditis elegans locomotion: Where do we stand? *Bioscience*, *64*(6), 476–486.

Gleeson, P., Lung, D., Grosu, R., Hasani, R., & Larson, S. D. (2018). c302: A multiscale framework for modelling the nervous system of Caenorhabditis elegans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1758), 20170379.

Hanakawa, T., Honda, M., Okada, T., Fukuyama, H., & Shibasaki, H. (2003). Neural correlates underlying mental calculation in abacus experts: A functional magnetic resonance imaging study. *NeuroImage*, *19*(2), 296–307. https://doi.org/10.1016/S1053-8119(03)00050-8

Henrich, J. (2016). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences*, *33*(2–3), 111. https://doi.org/10.1017/S0140525X10000725

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd Ed.). Sage.

Hornsby, J. (2012). *Ryle's knowing-how, and knowing how to act*. Oxford University Press. https://doi.org/oso/9780195389364.003.0003

Hutto, D. (2008). *Folk psychological narratives: The sociocultural basis of understanding reasons*. MIT Press.

Hutto, D. D., Gallagher, S., Ilundáin-Agurruza, J., & Hipólito, I. (2020). Culture in mind – An enactivist account: Not cognitive penetration but cultural permeation. In L. J. Kirmayer, S. Kitayama, R. Worthman, R. Lemelson, & C. A. Cummings (Eds.), *Culture, mind, and brain: Emerging concepts, models, applications* (pp. 163–187). Cambridge University Press.

Kaše, V., Hampejs, T., & Pospíšil, Z. (2018). Modeling cultural transmission of rituals in silico: The advantages and pitfalls of agent-based vs. System dynamics models. *Journal of Cognition and Culture*, 25. https://doi.org/10.1163/15685373-12340041

Kenrick, D. T., Keefe, R. C., Gabrielidis, C., & Cornelius, J. S. (1996). Adolescents' age preferences for dating partners: Support for an evolutionary model of life-history strategies. *Child Development*, *67*(4), 1499–1511. https://doi.org/10.2307/1131714

Kitano, H., Hamahashi, S., & Luke, S. (1998). The perfect c. elegans project: An initial report. *Artificial Life*, *4*(2), 141–156.

Kitayama, S., & Park, J. (2010). Cultural neuroscience of the self: Understanding the social grounding of the brain. *Social Cognitive and Affective Neuroscience*, *5*(2–3), 111–129. https://doi.org/10.1093/scan/nsq052

Kotseruba, I., & Tsotsos, J. K. (2018). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, 1–78.

Kronsted, C., Neemeh, Z. A., Kugele, S., & Franklin, S. (forthcoming). Modeling long-term intentions and narratives in autonomous agents. *Journal of Artificial Intelligence and Consciousness*.

Kugele, S., & Franklin, S. (2020). “Conscious” multi-modal perceptual learning for grounded simulation-based cognition. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 2459–2465). Cognitive Science Society.

Kugele, S., & Franklin, S. (2021). Learning in LIDA. *Cognitive Systems Research*, *66*, 176–200. https://doi.org/10.1016/j.cogsys.2020.11.001

Laland, K. N. (2017). *Darwin’s unfinished symphony: How culture made the human mind*. Princeton University Press.

Langton, C. G. (Ed.). (1997). *Artificial life: An overview*. MIT Press.

Latané, B., & Darley, J. M. (1969). Bystander “apathy.” *American Scientist*, *57*(2), 244–268.

Levin, D. T., & Banaji, M. R. (2006). Distortions in the perceived lightness of faces. *Journal of Experimental Psychology: General*, *135*(4), 501–512. https://doi.org/10.1037/0096-3445.135.4.501

Madl, T., Franklin, S., Chen, K., & Trappl, R. (2018). A computational cognitive framework of spatial memory in brains and robots. *Cognitive Systems Research*, *47*, 147–172. https://doi.org/10.1016/j.cogsys.2017.08.002

Mak, I. W., Evaniew, N., & Ghert, M. (2014). Lost in translation: Animal models and clinical trials in cancer treatment. *American Journal of Translational Research*, *6*(2), 114–118.

Malafouris, L. (2013). *How things shape the mind: A theory of material engagement*. MIT Press.

McAdams, D. (2006). *The redemptive self: Stories Americans live by*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195176933.001.0001

McCall, R., Franklin, S., Faghihi, U., Snaider, J., & Kugele, S. (2020). Artificial motivation for cognitive software agents. *Journal of Artificial General Intelligence*, *11*(1), 38–69.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

McLean, K. C., Boggs, S., Haraldsson, K., Lowe, A., Fordham, C., Byers, S., & Syed, M. (2020). Personal identity development in cultural context: The socialization of master narratives about the gendered life course. *International Journal of Behavioral Development*, *44*(2), 116–126. https://doi.org/10.1177/0165025419854150

Merleau-Ponty, M. (2012). *Phenomenology of perception* (D. A. Landes, Trans.). Routledge. (Original work published 1945)

Neemeh, Z. A., Kronsted, C., Kugele, S., & Franklin, S. (2021). Body schema in autonomous agents. *Journal of Artificial Intelligence and Consciousness*, *8*(1), 113–145. https://doi.org/10.1142/S2705078521500065

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. W. H. Freeman.

Nelson, K., & Fivush, R. (2020). The development of autobiographical memory, autobiographical narratives, and autobiographical consciousness. *Psychological Reports*, *123*(1), 71–96. https://doi.org/10.1177/0033294119852574

Newcomb, T. M. (1943). *Personality and social change; attitude formation in a student community*.

Newell, A. (1973). You can’t play 20 questions with nature and win: Projective comments on the papers of this symposium. In *Visual information processing*. Academic Press.

Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.

Odling-Smee, F. J., Laland, K. N., & Feldman, M. W. (2003). *Niche construction: The neglected process in evolution*. Princeton University Press.

Osorio, D., & Vorobyev, M. (1996). Colour vision as an adaptation to frugivory in primates. *Proceedings: Biological Sciences*, *263*(1370), 593–599. https://doi.org/10.1098/rspb.1996.0089

Overmann, K. A. (2017). Thinking materially: Cognition as extended and enacted. *Journal of Cognition & Culture*, *17*, 354–373. https://doi.org/10.1163/15685373-12340012

Overmann, K. A., & Wynn, T. (2019). On tools making minds: An archaeological perspective on human cognitive evolution. *Journal of Cognition and Culture*, *19*(1–2), 39–58. https://doi.org/10.1163/15685373-12340047

Petracca, E., & Gallagher, S. (2020). Economic cognitive institutions. *Journal of Institutional Economics*, 1–19. https://doi.org/10.1017/S1744137420000144

Pound, P., & Ritskes-Hoitinga, M. (2018). Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. *Journal of Translational Medicine*, *16*, 1–8.

Qin, J., & Wheeler, A. R. (2007). Maze exploration and learning in C. elegans. *Lab on a Chip*, *7*(2), 186–192.

Rankin, C. H. (2004). Invertebrate learning: What can’t a worm learn? *Current Biology*, *14*(15), R617–R618.

Ross, L., & Nisbett, R. E. (2011). *The person and the situation: Perspectives of social psychology*. Pinter & Martin Publishers.

Ryle, G. (1976). *The concept of mind*. Penguin Books.

Salska, I., Frederick, D. A., Pawlowski, B., Reilly, A. H., Laird, K. T., & Rudd, N. A. (2008). Conditional mate preferences: Factors influencing preferences for height. *Personality and Individual Differences*, *44*(1), 203–215. https://doi.org/10.1016/j.paid.2007.08.008

Sarma, G. P., Lee, C. W., Portegys, T., Ghayoomie, V., Jacobs, T., Alicea, B., Cantarelli, M., Currie, M., Gerkin, R. C., Gingell, S., Gleeson, P., Gordon, R., Hasani, R. M., Idili, G., Khayrulin, S., Lung, D., Palyanov, A., Watts, M., & Larson, S. D. (2018). OpenWorm: Overview and recent advances in integrative biological simulation of Caenorhabditis elegans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1758), 20170382. https://doi.org/10.1098/rstb.2017.0382

Schechtman, M. (1996). *The constitution of selves* (1st ed.). Cornell Univ. Press.

Schulte-Hostedde, A. I., Eys, M. A., & Johnson, K. (2008). Female mate choice is influenced by male sport participation. *Evolutionary Psychology*, *6*(1), 147470490800600113.

Schwarz, S., & Hassebrauck, M. (2012). Sex and age differences in mate-selection preferences. *Human Nature*, *23*(4), 447–466. https://doi.org/10.1007/s12110-012-9152-x

Seligman, R., Choudhury, S., & Kirmayer, L. J. (2016). Locating culture in the brain and in the world: From social categories to the ecology of mind. In J. Y. Chiao, S.-C. Li, R. Seligman, & R.

Turner (Eds.), *The Oxford Handbook of Cultural Neuroscience* (pp. 3–20). Oxford University Press.

Shackelford, T. K., Schmitt, D. P., & Buss, D. M. (2005). Universal dimensions of human mate preferences. *Personality and Individual Differences*, *39*(2), 447–458. https://doi.org/10.1016/j.paid.2005.01.023

Shults, F. L., & Wildman, W. J. (2018). Simulating religious entanglement and social investment in the Neolithic. In I. Hodder (Ed.), *Religion, history, and place in the origin of settled life* (pp. 33–63). University Press of Colorado. https://doi.org/10.2307/j.ctv3c0thf

Siegel, S. (2012). Cognitive penetrability and perceptual justification. *Noûs*, *46*(2), 201–222. https://doi.org/10.1111/j.1468-0068.2010.00786.x

Siegel, S. (2016). *The rationality of perception*. Oxford University Press.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv*. http://arxiv.org/abs/1712.01815

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Driessche, G. van den, Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), Article 7676. https://doi.org/10.1038/nature24270

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). MIT press.

Slaby, J., & Gallagher, S. (2014). Critical neuroscience and socially extended minds. *Theory, Culture & Society*, *32*(1), 33–59. https://doi.org/10.1177/0263276414551996

Sloman, A. (2001). Beyond shallow models of emotion. *Cognitive Processing*, *2*(1), 177–198.

Smuts, B. (1995). The evolutionary origins of patriarchy. *Human Nature*, *6*(1), 1–32.

Sterelny, K. (2003). *Thought in a hostile world: The evolution of human cognition*. Blackwell.

Stokes, D. (2013). Cognitive penetrability of perception. *Philosophy Compass*, *8*(7), 646–663.

Suzuki, M., Goto, T., Tsuji, T., & Ohtake, H. (2005). A dynamic body model of the nematode C. elegans with neural oscillators. *J. Robotics Mechatronics*, *17*(3), 318–326.

Szigeti, B., Gleeson, P., Vella, M., Khayrulin, S., Palyanov, A., Hokanson, J., Currie, M., Cantarelli, M., Idili, G., & Larson, S. (2014). OpenWorm: An open-science approach to modeling Caenorhabditis elegans. *Frontiers in Computational Neuroscience*, *8*, 137.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Uexküll, J. von. (2010). *A foray into the worlds of animals and humans with A theory of meaning* (J. D. O'Neil, Trans.). University of Minnesota Press.

Varela, F. J. (1988). Structural coupling and the origin of meaning in a simple cellular automation. In *The semiotics of cellular communication in the immune system* (pp. 151–161). Springer.

Varela, F. J., Thompson, E., & Rosch, E. (2016). *The embodied mind: Cognitive science and human experience* (Revised Ed.). MIT Press. (Original work published 1991)

Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W., Dudzik, A., Huang, A., Georgiev, P., Powell, R., Ewalds, T., Horgan, D., Kroiss, M., Danihelka, I., Agapiou, J., Oh, J., Dalibard, V., Choi, D., Sifre, L., … Silver, D. (2019, January 24). AlphaStar: Mastering the real-time strategy game StarCraft II. *DeepMind Blog*. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

Wade, T. J., Auer, G., & Roth, T. M. (2009). What is love: Further investigation of love acts. *Journal of Social, Evolutionary, and Cultural Psychology*, *3*(4), 290. https://doi.org/10.1037/h0099315

*Why are mice considered excellent models for humans?* (n.d.). The Jackson Laboratory. Retrieved September 14, 2020, from https://www.jax.org/why-the-mouse/excellent-models

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636. https://doi.org/10.3758/BF03196322

Wilson, S. W. (1991). The animat path to AI. In J.-A. Meyer & S. W. Wilson (Eds.), *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior* (pp. 15–21). MIT Press.