Simulations and the evolution of consciousness

Joshua Bensemann*, Padriac O'Leary, Yang Chen, Ludmila Miranda-Dukoski, and Michael Witbrock

University of Auckland

Abstract

We hypothesize that the emergence of consciousness in humans is directly related to the complexity, number of, and evolution of specialized cognitive systems. Here, we present our rationale and plan for an ongoing project to investigate the pathway to the emergence of consciousness via computer simulations of humans' evolutionary niche using artificial-intelligence agents. Agents will contain subsets of the specialized cognitive systems and will complete tasks modeled after pressures encountered by early humans. We will observe whether the increase in cognitive complexity, measured by the number and complexity of the specialized cognitive systems, leads to an increase in task performance.

Introduction

Consciousness has been the topic of speculation and analysis across multiple fields in the academic community. A drawback of the broad appeal that the study of consciousness has is that there is no consensus on the most basic concepts, including its definition. Some authors define a conscious agent as one that possesses a cognitive architecture with features such as memory, internal representations of the world, *et cetera* (Aleksander, 2007; Arrabales et al., 2010; Bensemann and Witbrock, 2021; Tononi and Koch, 2015). Proponents of Global Workshop Theory (Baars, 2005) go further along this route, suggesting that consciousness is itself a system built from interactions between highly-specialized cognitive systems (e.g., attention, memory, etcetera.). This definition implies that consciousness is an emergent property of highly-specialized cognitive systems (Zlomuzica and Dere, 2022). In other words, the evolution of consciousness in *Hominini* (i.e., modern humans and their ancestors) is taken to be directly related to the evolution of these specialized systems.

Our project focuses on investigating the emergence of consciousness in humans by replicating the conditions under which *Hominini* evolved. We will be defining consciousness using a third-person introspective model (Choifer, 2018). The specialized system central to this definition is the Theory of Mind (ToM; Premack and Woodruff, 1978). ToM is the capacity to create models that are used to predict the knowledge and motivations of others. Once an agent can model the minds of others, that agent can also apply that capability to itself, resulting in a model of its own mind.

We hypothesize that various evolutionary pressures lead to increases in complexity as well as the interconnectedness between various specialized cognitive systems such as attention, communication, and, critically, ToM. In order to methodically examine the characteristics of the set of interconnected specialized cognitive systems required to produce given levels of cognitive complexity along with the ecological niches that lead to these characteristics, we will build digital environments and agents using computer simulations and artificial intelligence (AI). This approach is known as synthetic ethology (MacLennan, 2007).

The core of our research strategy is the creation of a computer-simulated environment based on early hominin ecology. However, allowing computer scientists to create environments for testing can introduce an implicit bias towards producing desired results by only incorporating what they consider important (Laird, 2001). To minimize this, we are using current models of evolutionary ecology to guide the development of the simulation. Our initial approach will build on the human evolution models of Kim Sterelney, who views our cognitive evolution as based on the gradual progression of interaction between individual and social feedback loops (Sterelny, 2012; Sterelny et al., 2013).

Our work on creating artificial replicas of early hominin environments is in progress. Once completed, we will introduce AI agents. Various agents controlled by different cognitive models of varying complexity will be tested. Controlling both the composition of the environment and the agent's cognitive makeup will enable us to experimentally identify any advantages that various sets of specialized cognitive systems afford us. In essence, we are performing a step-wise regression to generate enough data to analyze the characteristics of the set of specialized cognitive systems that might underpin human consciousness while accounting for factors such as model complexity.

*Email: josh.bensemann@auckland.ac.nz

Conscious Agents

Our AI agents will have various subsets of the specialized cognitive systems whose development is thought to correlate to the emergence of consciousness. The agents' goal will be to complete tasks thought to have been encountered by early humans. We will observe whether cognitive complexity leads to improvements in task performance.

Our approach to building artificial consciousness is similar to those suggested by others in prior work (e.g., Aleksander, 2007; Horton et al. 2013). One such approach is the ConsScale (Arrabales et al., 2010), an ordinal scale developed to incorporate other consciousness models. To reach each level in the scale, the agent must possess a minimum set of architectural and behavioral features from all previous levels, with levels ordered based on the likely phylogenetic path to our species' consciousness. Pre-existing plans such as the ConsScale will be used as starting point for our agents. However, the development will be guided by the work of others to ensure its cognitive development is consistent with our working definition of consciousness.

We turned to evolutionary psychology to make an educated guess about the highly-specialized cognitive systems that we must be included. Mounting evidence from the field suggests that sociality and solving other niche-specific problems played a significant, if not pivotal, role in the evolution of human consciousness and cognition. Of the number of systems involved in human socialization, ToM is of particular interest to us and the current research.

The first generation of our agents will be built using components from the various existing computational ToM models. These core components include beliefs, desires, and memory (or functional equivalents). We will inject ToM into AI agents and provide them with basic knowledge of the conditions of their virtual environment and an array of preferences for manipulating the contents of the environment. Our ultimate aim is to build an agent capable of cooperative behavior and formulate causal stories about their environments and other agents.

We know from the literature that various versions of cognitive capacities will produce varied results. For example, variations of ToM components can have noticeable effects on an agent's performance when learning to compete or cooperate with other agents. Experiments with the recursive aspects of ToM have shown that agents who could model deeper levels of recursion increased group performance when cooperating and bested competitors who possessed lesser recursive capabilities (Devaine et al., 2014). Similarly, having a more extended memory of the past actions of other agents allows an agent to better compete or cooperate with other agents (Anh et al., 2011). However, increasing memory length increases the cognitive cost, which might outweigh the gain in performance (Han et al., 2012).

The Environment

Deriving the environment that led to the emergence of consciousness is a critical requirement in understanding the conditions that gave rise to consciousness. By recreating the environment that human ancestors evolved in digitally, other researchers have uncovered conditions that may have led to the evolution of communication (Gong and Shuai, 2013; MacLennan, 2007; Miikkulainen and Li, 2016). For example, Miikkulainen and Li created a "jungle world" where pairs of agents' fitness increased if both agents chose to hunt or both chose to mate. They demonstrated that the evolution of communication was unnecessary when the simulation was full-observable to both agents. Communication became necessary when the environment was partially-observable and increased fitness when shared information was needed for cooperation. While our simulation will not necessarily require evolution, the principle is the same; our environment will be designed to test whether cognitive components provide any fitness advantage to agents that possess them. This allows a theoretical test for the utility of consciousness by adding its theoretical precursors to the agents.

Our core design principle for the environment is replicating resource-gathering and predation problems from our evolutionary niche. There will be multiple agents within an environment, each having partial information about the world. An agent's ability to forage individually and in groups will affect their fitness. Group activities include hunting or division of labor to increase overall fitness. By having both group and independent tasks, the environment will test fitness values of agents when acting alone compared to acting within a group.

The first version of the environment will be a 2D turn-based grid word. These settings were chosen to begin testing our agents as quickly and efficiently as possible as our agents will not require complex visual recognition systems as in the agents tested by the animal-AI testbed (Crosby et al., 2020). By creating a grid-based system, we can provide the agents with a simpler world representation. However, we will implement various grids - each containing information from a different sensory modality - to increase environmental complexity.

Multiple metrics such as survival time, resources gathered, and others will be used to measure evolutionary fitness. We will also measure cognitive efficiency by comparing the cognitive complexity of the agents to their task scores. Cognitive complexity will be measured using metrics such as the number of operations required for decisions, memory size, learning rates, or any commonly used metric in either psychology or the computer sciences. Suppose we normalize any survival metrics we use concerning an agent's cognitive complexity and show that an agent's gains in survival exceed the price. In that case, we have evidence that such cognitive capacities evolved due to their significant advantage.

References

Aleksander, I. (2007). Why axiomatic models of being conscious? *Journal of Consciousness Studies*, 14(7):15–27.

Anh, H. T., Moniz Pereira, L., and Santos, F. C. (2011). Intention recognition promotes the emergence of cooperation. *Adaptive Behavior*, 19(4):264–279.

Arrabales, R., Ledezma, A., and Sanchis, A. (2010). Consscale: A pragmatic scale for measuring the level of consciousness in artificial agents. *Journal of Consciousness Studies*, 17(3-4):131–164.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in brain research*, 150:45–53.

Bensemann, J. and Witbrock, M. (2021). The effects of implementing phenomenology in a deep neural network. *Heliyon*, 7(6):e07246.

Choifer, A. (2018). A new understanding of the first-person and third-person perspectives. *Philosophical Papers*, 47(3):333–371.

Crosby, M., Beyret, B., Shanahan, M., Hernández-Orallo, J., Cheke, L., and Halina, M. (2020). The animal-ai testbed and competition. In *NeurIPS 2019 competition and demonstration track*, pages 164–176. PMLR.

Devaine, M., Hollard, G., and Daunizeau, J. (2014). Theory of mind: did evolution fool us? *PloS One*, 9(2):e87619.

Gong, T. and Shuai, L. (2013). Computer simulation as a scientific approach in evolutionary linguistics. *Language Sciences*, 40:12–23.

Han, T. A., Pereira, L. M., and Santos, F. C. (2012). Corpus-based intention recognition in cooperation dilemmas. *Artificial Life*, 18(4):365–383.

Horton, J. D., Francis, M., and Sußenburger, E. (2013). A long term proposal to simulate consciousness in artificial life. In *ICAART (2)*, pages 389–394.

Laird, J. E. (2001). Using a computer game to develop advanced ai. *Computer*, 34(7):70–75.

MacLennan, B. (2007). Making meaning in computers: Synthetic ethology revisited. In *Artificial Cognition Systems*, pages 252–283. IGI Global.

Miikkulainen, R. and Li, X. (2016). Evolving artificial language through evolutionary reinforcement learning. In *ALIFE 2016, the Fifteenth International Conference on the Synthesis and Simulation of Living Systems*, pages 484–491. MIT Press.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Sterelny, K. (2012). *The evolved apprentice*. MIT press.

Sterelny, K., Joyce, R., Calcott, B., and Fraser, B. (2013). *Cooperation and its evolution*. MIT Press.

Tononi, G. and Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668):20140167.

Zlomuzica, A. and Dere, E. (2022). Towards an animal model of consciousness based on the platform theory. *Behavioural brain research*, 419:113695.