Synthesizing Sentience: Integrating Large Language Models and Autonomous Agents for Emulating Human Cognitive Complexity

Jay Ratican  
*Lindenwood University*, jratican@lindenwood.edu

James Hutson  
*Lindenwood University*, jhutson@lindenwood.edu

Daniel Plate  
*Lindenwood University*, dplate@lindenwood.edu

Follow this and additional works at: https://digitalcommons.lindenwood.edu/faculty-research-papers

Part of the Artificial Intelligence and Robotics Commons

Recommended Citation  
https://digitalcommons.lindenwood.edu/faculty-research-papers/513

This Article is brought to you for free and open access by the Research and Scholarship at Digital Commons@Lindenwood University. It has been accepted for inclusion in Faculty Scholarship by an authorized administrator of Digital Commons@Lindenwood University. For more information, please contact phuffman@lindenwood.edu.
Synthesizing Sentience: Integrating Large Language Models and Autonomous Agents for Emulating Human Cognitive Complexity

Jeremiah Ratican*, James Hutson, Daniel Plate

Lindenwood University, USA


Received: 02 October, 2023; Accepted: 10 October, 2023; Published: 20 October, 2023

*Corresponding author: Jeremiah Ratican, Lindenwood University, USA. Email: jratican@lindenwood.edu

Copyright: © 2023 Ratican J., et al., This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

A B S T R A C T

The paper aims to present a novel methodology for emulating the intricacies of human cognitive complexity by ingeniously integrating large language models with autonomous agents. Grounded in the theoretical framework of the modular mind theory—originally espoused by Fodor and later refined by scholars such as Joanna Bryson—the study seeks to venture into the untapped potential of large language models and autonomous agents in mirroring human cognition. Recent advancements in artificial intelligence, exemplified by the inception of autonomous agents like Age in GPT, auto GPT, and baby AGI, underscore the transformative capacities of these technologies in diverse applications. Moreover, empirical studies have substantiated that persona-driven autonomous agents manifest enhanced efficacy and nuanced performance, mimicking the intricate dynamics of human interactions. The paper postulates a theoretical framework incorporating persona-driven modules that emulate psychological functions integral to general cognitive processes. This framework advocates for the deployment of a plurality of autonomous agents, each informed by specific large language models, to act as surrogates for different cognitive functionalities. Neurological evidence is invoked to bolster the theoretical architecture, delineating how autonomous agents can serve as efficacious proxies for modular cognitive centers within the human brain. Given this foundation, a theory of mind predicated upon modular constructs offers a fertile landscape for further empirical investigations and technological innovations.

Keywords: Autonomous Agents, Large Language Models, Cognitive Complexity, Modular Mind Theory, Theory of Mind

1 Introduction

The rapid maturation of generative artificial intelligence (AI) technologies and large language models (LLMs) has ushered the academic community into a new era of possibilities and challenges. Recent advancements, such as OpenAI’s ChatGPT 3.5 and 4, have demonstrated unparalleled aptitude in generating linguistically rich and contextually coherent text, provoking intricate dialogues that closely resemble human-to-human interaction (Gill & Kaur, 2023). Concurrently, these evolutions have imbued the intellectual discourse with pressing questions about the future trajectory of AI, particularly within a myriad of application domains such as healthcare, education, and assisted living (Albahri et al., 2023; Lee, 2023; Liu et al., 2023; Shahriar Hayawi, 2023).

In light of these developments, the endeavor to understand and replicate human cognitive complexity has gained significant traction within the academic landscape. Such pursuits inevitably broach the compelling research frontiers of integrating human emotional dimensions into AI paradigms (Latif et al., 2023). Previous studies have documented that the machine learning algorithms underpinning LLMs are capable of simulating and adopting a diverse set of “personali(13) ty,” thereby laying the groundwork for the exploration of artificial emotional dimensions (Garon, 2023). These technological strides bring to the fore the tantalizing possibility of constructing emotionally intelligent AI systems, which hold the promise of fostering enhanced user experiences through more meaningful interactions (Ray, 2023).

Central to the arguments presented herein is the integration of large language models with autonomous agents. The impetus for this amalgamation is derived from modular mind theory—a foundational concept which posits that cognitive processes are compartmentalized within the brain into distinctmodules...
This theoretical underpinning has been further refined by scholars, most notably Joanna Bryson, who contend that modularity exists not only in natural cognitive systems but can also be synthesized within artificial constructs (Bryson, 2010).

Building upon this rich theoretical heritage, as well as a burgeoning body of empirical research such as the work of Park et al. (2023), which highlighted the potential of autonomous agents to mimic human behavior when equipped with personal motivations and preferences this paper endeavors to present a novel strategy. Specifically, it posits an innovative approach for emulating human cognitive complexity through the integration of large language models and autonomous agents. By doing so, the study aims to make a substantive contribution to the extant literature, while also laying a fertile groundwork for future empirical investigations and technological innovations.

The amalgamation of large language models with autonomous agents, structured upon the theoretical scaffolding of modular mind theory, signifies a seminal contribution to interdisciplinary research within the realms of artificial intelligence, psychology, and neuroscience. Leveraging cutting-edge advancements in machine learning architectures and autonomous system designs, the proposed model endeavors to construct a synthetic analog of human cognition, partitioned into specialized modules corresponding to identified facets of mental processing. Grounded in robust empirical observations and theoretical perspectives from the partitioning of the human mind as per modular theories to the discernible efficacy of persona-driven autonomous agents the model introduces an innovative methodological approach for simulating human cognitive complexity. In doing so, it not only pushes the boundaries of what artificial systems can achieve in terms of mimetic cognitive functions, but also lays the groundwork for further research that might elucidate the enigmatic intricacies of human cognition. Thus, the model holds considerable promise both as a heuristic tool for cognitive science inquiries and as a technological harbinger for increasingly sophisticated artificial intelligence systems.

2. Literature Review

2.1 AI Agents: Emotions and Personas

Scholarly inquiries into emotionally intelligent artificial intelligence (AI) have proliferated over recent years, revealing multifaceted dimensions that encompass not only the technological underpinnings but also the psychosocial and ethical implications. The collective body of research underscores the pivotal role of emotional faculties in intelligent behavior and decision-making processes (Duan et al., 2019; Mahmud et al., 2022; Strich et al., 2021). Mahmud et al. (2022), for instance, argue that emotions serve as a complex signaling system that affects intelligent decision-making, thereby accentuating the integral nature of emotions in cognitive functions. The importation of these emotional aspects into artificial cognitive systems constitutes a seminal shift in the landscape of AI development (Zall & Kangavari, 2022).

Frameworks for AI systems imbued with social and emotional capabilities have been proposed as necessary advancements to facilitate nuanced human-machine interactions (Samsonovich, 2020; Picard et al., 2004). Samsonovich (2020) astutely observes that empathy in artificial agents contributes to more authentic social interactions, thereby substantiating the claim that emotional intelligence in AI can augment the quality of social exchanges. These proposals intersect with ongoing efforts to construct AI-driven communication systems that employ affective computing techniques to decipher and simulate emotional states (Li et al., 2019; Khachane, 2017). Such research culminates in practical applications that range from emotionally intelligent chatbots to interfaces designed for specialized sectors.

On the frontiers of ethics and practicality, scholars have broached the challenges and conundrums that accompany the incorporation of artificial emotions into AI systems (Pusztahelyi, 2020; Cominelli et al., 2021). Pusztahelyi (2020) raises a pertinent question: What ethical considerations arise when machines simulate human emotions? The query points to a nascent but critical discourse on the moral boundaries and responsibilities involved in deploying emotionally intelligent AI. Concurrently, the realm of affective computing has commenced a transitional phase towards more robust cognitive models of emotional intelligence (Li et al., 2022; Wortman & Wang, 2022), thereby indicating a persistent interest in enhancing the efficacy and reliability of these technologies.

Healthcare, among other sectors, is posited as a fertile ground for the deployment of emotionally intelligent AI systems, particularly in the context of emotion recognition tools that could assist medical professionals (Marcos et al., 2021). The discussion extends to the ethical landscape, probing the extent to which such AI systems should be vested with autonomous decision-making capacities (Huh & Seo, 2019). Additionally, Andersson (2022) foregrounds the necessity of contemplating the potential infringement on fundamental human rights, such as freedom of thought, when deploying emotionally intelligent AI.

Thus, the existing literature serves as an intellectual tapestry woven with diverse threads of academic interest: from the rudimentary mechanics of emotional intelligence to its practical applications and ethical ramifications. Research has delved into the role of emotions in intelligent behavior, articulated the benefits of integrating emotional aspects into artificial cognitive systems, and offered frameworks to realize socially and emotionally intelligent AI. Concurrently, inquiries into AI-enabled communication mechanisms, advancements in synthetic emotional intelligence, and the ethical contours of artificial emotions have enriched the academic discourse. Finally, considerations of healthcare applications and legal challenges offer a holistic view of the state and prospects of emotionally intelligent AI.

2.2 Generative AI Agents

The incorporation of generative AI agents into the domains of video gaming and decision-making processes manifests as a burgeoning area of scholarly inquiry. Early studies, such as those conducted by Naddaf (2010) and Liu et al. (2017), have explored the implementation of reinforcement learning-based methods to tutor AI agents in gameplay. Naddaf’s work provides foundational understanding of how AI agents can be trained through reinforcement learning to adapt to specific game environments. Liu et al. (2017) further evolved these concepts, employing agents like the Random Mutation Hill-Climber to bring about game versions with substantial skill depth. These studies are complemented by the work of Holmgård et al. (2014) and Barthet et al. (2022), which postulate that artificial agents can function as abstract simulations of human players’ internal decision-making processes. Barthet et al. note that generative personas manifest behaviors and responses that closely emulate human personas, affirming the potential for AI agents to accurately model human behavior and decision-making.
A related vein of research has delved into the evaluation and testing of video games via AI agents. For instance, Ariyurek, Betin-Can, & Surer (2019) found that synthetic agents demonstrate a capacity comparable to human testers in identifying software glitches. However, Fathi & Palhang (2018) delineate a notable limitation, pointing out the paucity of diversity in agent behavior, leading to predictability. The work of Nareyek (2000), Tan and Nareyek (2009), and Miikkulainen et al. (2006) subsequently elucidate the wide-ranging potential of AI techniques in modern video gaming, while Fernández et al. (2006) dissect the complexities involved in designing the behavior of automated player characters.

Emerging research has also scrutinized the capabilities of generative AI models such as ChatGPT in reshaping practices across diverse scientific and medical fields. Morris (2023) reveals through interviews with twenty scientists that generative AI holds the potential to accelerate the pace of scientific discovery. Furthermore, Megahed et al. (2023) indicate that these models could enhance statistical process control practices. They do, however, caution against potential misuse and misunderstanding, given the nascent stage of these technologies. Contributions from Murphy & Thomas (2023) explore the deployment of generative AI in spinal cord injury research, illuminating its utility in creating virtual models and optimizing medical protocols.

Recent advancements demonstrate a unique capability of generative AI agents to simulate diverse persons in controlled environments. Research conducted by Park et al. (2023) manifests this through the creation of an RPG-style virtual world populated by AI agents with distinct personalities and social dynamics. The investigators employed the ChatGPT API for social interactions, developing an intricate architecture that simulated agents with both memories and experiences. When evaluated for the believability of behavior, the generative agent architecture surpassed even human role-play responses in terms of authenticity. However, the researchers also sounded notes of caution, warning of ethical considerations including the formation of inappropriate parasocial relationships and an overreliance on generative agents.

Taken together, scholarship illuminates a trajectory wherein generative AI agents are increasingly implicated in an array of contexts ranging from video gaming to scientific research and medical applications. While the capabilities of these agents to mimic human behavior and decision-making have been affirmed, caveats regarding their responsible deployment remain. Ethical considerations, particularly in the context of believability and human-AI interaction, require further scholarly exploration to safeguard against unintended consequences. Therefore, the research corpus underscores the transformative potential of generative AI agents, even as it calls for a nuanced understanding of their limitations and ethical implications.

2.3 The Modular Mind Theory and Its Application

The Modular Mind Theory postulates that the human brain operates not as a monolithic entity, but as an intricate network of specialized modules or centers dedicated to specific cognitive functions. This perspective gains empirical substantiation through biological evidence, most compellingly through cases of traumatic brain injuries. Observations of individuals who have incurred such injuries often reveal precise deficits corresponding to the localized areas of damage. For instance, damage to Broca’s area has been found to result in specific language impairments without affecting other cognitive faculties (Damasio & Damasio, 1992). Fodor (1983) advanced theories around modularity of mind, arguing for specialized and autonomous subsystems that operate on distinct types of information. Furthermore, Fodor posited that such modules are domain-specific, operate autonomously, and are evolutionarily hardwired.

The construct of the modular mind gains additional validation through groundbreaking split-brain experiments. Gazzaniga (1970) pioneered work that involved severing the corpus callosum, the band of nerve fibers that connects the two hemispheres of the brain. The resultant observations illuminated how each hemisphere functions as a distinct module capable of independent cognitive processes. Psychoanalysts, including Fodor, have drawn on these findings to further elaborate the modularity theory, contributing to a deeper understanding of cognitive architecture.

Extending the theory of the modular mind into the realm of AI, Bryson (2000) offers the seminal treatment “Modular Representations of Cognitive Phenomena in AI, Psychology, and Neuroscience.” This work builds upon the notion that the architecture underlying cognitive processes whether biological or artificial is fundamentally modular in nature. Bryson delineates a comprehensive framework that distinguishes between horizontal and vertical modules. According to Bryson, horizontal modules pertain to general cognitive processes like attention and memory, whereas vertical modules are domain-specific and tailored for particular tasks. Moreover while the modular perspective affords a parsimonious representation of complex cognitive phenomena, Bryson emphasizes the manifold challenges in constructing modular representations that capture the nuanced interplay between various cognitive components. The task becomes even more daunting when one considers the need to integrate findings across multiple disciplines psychology, neuroscience, and AI to form a cohesive and comprehensive model.

The intricate nature of these systems whether a neural network in a biological entity or a machine learning model in AI reveals a complex tapestry of interwoven modules. The challenge for contemporary research lies in mapping these modular structures accurately and comprehensively, a task made more complicated by the labyrinthine interdependencies that exist within and across modules. Therefore, as Bryson (2000) points out, while modularity offers a framework for simplifying the complexity inherent in cognitive systems, it simultaneously highlights the intricate and multi-faceted nature of these systems.

Thus, the modular mind theory offers an explanatory framework that enjoys empirical support from multiple disciplines. From the selective impairments observed in traumatic brain injuries to the domain-specific architectures posited in artificial intelligence, the modular perspective offers compelling insights. Yet, as research progresses, it becomes evident that capturing the complexity of these modular systems, whether in the human brain or artificial entities, remains a formidable intellectual challenge.

3. Recommendations

3.1 The Proposed Synthesis of Sentience

Contemporary large language models, such as ChatGPT and its more advanced counterpart GPT-4, represent an architecture underpinned by sophisticated neural networks trained on copious amounts of textual data. These architectures demonstrate remarkable flexibility and adaptability, particularly when guided by precise prompts or augmented by additional computational
components (Elshentenawy et al., 2023). The capabilities range from rudimentary conversational tasks to solving complex analytical problems, thereby embodying traits suggestive of multiple cognitive modules within a unified framework.

Recent scholarship posits the integration of LLMs with autonomous agents as a pathway toward emulating the complexity inherent in human cognition. In essence, autonomous agents would represent specific psychological modules, thus functioning as surrogates for distinct components of human mental activity (Hong et al., 2023; Xi et al., 2023). A compelling aspect of this synthesis is the incorporation of personas, delineated behavioral and cognitive profiles, to guide the activities of these agents (Ratican & Hutson, 2023). The overarching objective centers on mimicking the variegated complexity of human thought, achieved through orchestrating the interactions among these autonomous modules.

Essential to this synthetic model of sentience is the postulation of an executive module, responsible for the preservation of a cohesive sense of self. An executive module would function analogously to the role played by the prefrontal cortex in human cognition, overseeing the activity of specialized modules such as "emotion," "social cognition," "executive control," and "cognitive processing" (Williams et al., 2023). This structure aims to replicate the nuanced interplay observed in human cognitive systems, where specialized regions of the brain handle specific tasks yet work in concert to produce unified thought and behavior.

In fact, the empirical basis for such a synthesis is not without precedent. Cases of traumatic brain injury have demonstrated how specific cognitive functions are impaired in direct correlation to the location of damage, thereby reinforcing the modular view of the mind (Damasio & Damasio, 1992). Furthermore, research into split-brain patients highlights the semi-autonomous nature of brain hemispheres (Gazzaniga, 1970). Theories of modularity from cognitive science and psychoanalysis further lend credence to this approach (Fodor, 1983). However, the complexity of replicating the architecture of human cognition should not be underestimated. Efficacy depends on precise calibration of each autonomous module’s functional parameters and effective integration into an overarching system. Furthermore, the algorithmic architecture must continually adapt to account for the emergent properties of the integrated system, an area necessitating further investigation.

The proposed synthesis offers a theoretical yet empirically grounded framework for advancing the capabilities of large language models. By integrating these models with specialized autonomous agents, each embodying a distinct cognitive or psychological module, this architecture aims to approximate the complexity and functionality of human cognition. Although these are early days for such interdisciplinary ventures, the approach stands as an exemplar of converging insights from psychology, neuroscience, and artificial intelligence, holding promise for both theoretical advancement and practical application.

3.2 The Proposed Model: Autonomous Agents as Proxies for Mental Modules

The conceptual bedrock of the present study builds upon extant literature that delineates the human mind as a modular entity a concept espoused by theorists like Fodor (1983) and empirical researchers alike (Bryson, 2005). Specifically, Bryson’s seminal paper serves as a lynchpin, situating this study at the crossroads of neuroscience, psychology, and AI. Bryson delineates the mind as comprising both horizontal and vertical modules, where the former correspond to generalized cognitive processes and the latter to domain-specific skills. This dual-module model offers a comprehensive approach to cognitive phenomena, thus laying the groundwork for the current research initiative.

Current AI models such as GPT-4 consist of neural networks trained on vast datasets, demonstrating adaptability and functional versatility (Brown et al., 2020). Such versatility lends credence to the feasibility of employing these models as proxies for psychological and neurological modules, as delineated by modular mind theory and related psychological theories. For instance, Brown et al. (2020) point out the ability of these language models to adapt to various task requirements, a trait that could serve well in emulating the flexibility of human cognition. Central to the proposed model is the idea of employing autonomous agents as representatives for specific mental modules. Each agent, powered by large language models potentially of varying capacities is accorded a specialized role reflective of specific mental functions. These agents then process information and relay it to an ‘executive module,’ thereby preserving an overarching sense of self. Herein, the model takes inspiration from the biological observations that trauma to specific brain regions often results in correlated cognitive impairments (Norman et al., 2023).

The issue of computational efficiency is pivotal; thus, the model postulates the potential utility of smaller, specialized models for specific modules. For example, less computationally demanding pre-trained models, such as BERT-Tiny and BERT-Small, if fine-tuned appropriately, could outperform a larger, general-purpose model like GPT-4 in specialized tasks (Rana et al., 2023). The crux of the model resides in the executive function module, a computational entity that ensures mental cohesion and preserves a unified sense of self. Biologically speaking, the human brain appears to possess an analogous function, particularly within its hemispheres (Gazzaniga, 1983). Gazzaniga’s work on split-brain patients (and more recently Zhu (2023)) reveals that the two hemispheres can hold divergent beliefs yet maintain a unified self, suggesting an executive function at play. Challenges notwithstanding, this theoretical framework offers a robust scaffold for future research. The proposition of using autonomous agents as proxies for mental modules melds advances in AI with nuanced psychological theories. The ultimate aim achieving a coherent, unified agent mirrors the intricate interplay of modular functions within the human mind, thereby extending the frontiers of both AI and cognitive science.

4. Conclusion

The endeavor to integrate autonomous agents as representative elements of distinct psychological and neural functions presents a research frontier characterized by both conceptual promise and empirical challenges. A noteworthy aspect demanding scholarly attention is the identification of optimal architectures for these agents, with the imperative of ensuring compatibility with established theories of mental modules. Experiments are requisite to ascertain the computational frameworks best suited for the emulation of specific cognitive, affective, and sensorimotor functions, thereby fostering an alignment between psychological understanding and computational realization.

Efficiency remains a paramount consideration in the allocation of roles among different artificial intelligence
agents. The diversity among available large language models, ranging from those tailored for general purposes like GPT-4 to specialized variants like TinyBERT, provides researchers with a broad spectrum of options. Indeed, the application of smaller, specialized models like TinyBERT could enhance the system’s efficiency without substantially compromising functionality. TinyBERT, for instance, has been observed to perform effectively in specific tasks while requiring less computational power, rendering it an attractive candidate for specialized roles within the synthetic mental framework (He et al., 2023).

Central to the success of this synthetic cognitive model is the maintenance of a cohesive and unified mental state, simulated through an executive function module. This feature not only parallels the integrative operations within human cognition but is also essential for the functional stability of the model. It ensures that the individual components coalesce into an entity approximating human-like cognition, sidestepping the potential for conflicting outputs that might compromise the system’s integrity. The interdisciplinary orientation of this research engaging computational science, neuroscience, and psychology is instrumental in providing a multifaceted understanding of cognition. By employing modular representations, this work navigates the complexities of mental functions and creates a scaffold upon which advancements in individual disciplines can be collectively integrated.

Finally, the potential implications of this research are manifold, notably in the domain of emulating intricate facets of human cognition. Although the endeavor remains exploratory, the plausible outcomes could significantly expand the scope of what artificial intelligence systems can achieve. Nevertheless, the synthetic cognition model presented herein accentuates the parallels as well as the distinctions between biological and artificial cognitive architectures. While the modular approach has gained traction in the understanding of both natural and artificial cognitive systems, it is pivotal to remember that the latter still lacks the biological nuances that characterize the human mind, such as neuroplasticity and emotional complexity. In all, the quest for a synthetic model that approximates the multifaceted nature of human cognition posits not only scientific challenges but also opportunities for revolutionary advancements in both artificial intelligence and neuroscience.

5. Data Availability
Data available upon request.

6. Conflicts of Interest
The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

7. Funding Statement
NA

8. Authors’ Contributions
Conceptualization, J. Ratican; Methodology, J. Ratican; Validation, J. Hutson; Investigation, J. Hutson – Original Draft Preparation, J. Hutson; Writing – Review & Editing, J. Hutson.; Visualization, J. Hutson.

9. About The Authors
James Hutson: Department Head of Art History and Visual Culture, College of Arts and Humanities, Lindenwood University, Saint Charles, MO, USA.

10. References


22. Foster NF, & Gibbons S. Understanding Faculty to Improve Content Recruitment for Institutional Repositories. Online Submission 2005;11(1).


51. Megahed FM, Chen YJ, Ferris JA, Knoth S, & Jones-Farmer LA. How Generative AI Models such as ChatGPT can be (Mis)Used in Practice, Education, and Research? An Exploratory Study. Quality Engineering 2023;1-29.


