Computational Consciousness for System-2 AI: Theory Survey and Model Design

Edward Y. Chang
Stanford University
echang@cs.stanford.edu
January 7^{th}, 2023 (Version 1.0)

***Abstract*—This paper aims to create a model of consciousness for system-2 AI, which can handle tasks involving reasoning, planning, emotion, and decision making. We examine principles from philosophers and theories from psychiatrists and neuroscientists based on observations from various empirical studies. While the science community's understanding of consciousness and its location in the brain is incomplete, we believe that the implementation of consciousness in machines does not have to strictly follow human anatomy. We employ a functionalism approach to design the Computational Consciousness Model (CCM). CCM consists of four modules and three subsystems, which utilize scheduling and reinforcement learning algorithms. To tailor the reward systems to individuals and their local cultures and laws, we suggest using prompting templates and soliciting user feedback. We demonstrate that CCM is capable of supporting the functionalities required of an attentive, loving, and empathetic long-term care agent.**

I. Introduction

Narrow AI, also known as system-1 AI, is designed to perform specific, pre-defined tasks efficiently using machine learning algorithms. Examples of system-1 AI include image recognition and language translation systems. While effective at the narrow tasks they are designed for, these AI systems are limited in their ability to perform more complex tasks such as reasoning, planning, emotion interpretation, and decision-making.

To address these limitations, researchers have proposed developing system-2 AI, which is inspired by the human ability to perform more complex cognitive functions such as understanding and interpreting information and making logical decisions. The relationship between system-2 and system-1 AI can be compared to the relationship between human consciousness and unconsciousness, with system-1 AI being similar to the unconscious mind that performs automatic and reflexive actions, and system-2 AI aiming to replicate some of the capabilities of human consciousness that allow for more complex thought processes.

What is consciousness? Consciousness is the awareness of one's own mental states and experiences, and is characterized by subjective experiences such as thoughts, feelings, and sensations. It is a complex and multifaceted concept that has been studied for centuries by philosophers, scientists, and theologians, but the precise nature of consciousness and how it arises from the brain and other biological systems is still not fully understood. Some theories propose that consciousness is a fundamental property of the universe (panpsychism), while others suggest that it emerges from complex computations in the brain (functionalism). There are several theories about the nature of consciousness and how it arises, including the global workspace theory [Baars(1988)], the integrated information theory [Tononi(2004)], [Tononi(2008)], [Tononi(2016)], the neural correlates of consciousness approach [Crick and Koch(2003)], [Koch and Tsuchiya(2012)], and attention schema theory [Graziano(2013)], [Graziano(2016)]. These theories offer different perspectives on the underlying mechanisms of consciousness and the roles of specific brain areas and neural processes in generating subjective experience. Empirical studies of consciousness have also contributed to our understanding of the phenomenon. For example, studies on altered states of consciousness, such as those induced by sleep, meditation, or psychoactive drugs, have provided insight into how the brain's activity patterns change in these states and how they relate to changes in subjective experience. We survey and discuss theories of consciousness in Section II.

To model and develop a system that exhibits human-like consciousness or system 2, it is important to define the goals and desired functionalities of such a system. In Section III, we employ the approach of functionalism to define a list of capabilities that we consider system-2 AI should support. We later validate the list through a case study of an intelligent healthcare agent, which is demanded to be knowledgeable, attentive, loving, empathetic, humorous, patient, and adaptive to individual needs and preferences. Based on the functional specifications, we propose our Computational Consciousness Model (CCM) to address these functions. The CCM includes modules, subsystems, and algorithms to enable state transitions, priority-based scheduling, and reward-based optimization. The details of the CCM are depicted in Section IV.

In Section V, we use a case study of a long-term care agent to evaluate the performance of the CCM. The case study is designed to test the capabilities of the CCM in a realistic and practical scenario, and to demonstrate how the CCM can support the required functions of system-2 AI. In the case study, the agent is required to provide knowledgeable, attentive, and loving care to an elderly patient over a long period of time. The agent must be able to adapt to the individual needs and preferences of the patient, and to maintain a high level of empathy and emotional intelligence.

Finally, in Section VI, we offer concluding remarks and discuss open issues and future research directions for the development of system-2 AI. We believe that the CCM is

a promising approach for modeling and developing advanced AI systems that exhibit human-like consciousness, and that it has the potential to contribute to the development of more intelligent, empathetic, and adaptive AI systems that can benefit society in various ways.

II. What is Consciousness

In order to overcome the limitations of narrow AI (e.g., lacking interpretability, robustness, and generalization), researchers (e.g., [Bengio(2020)], [Chang(2020-22)]) have proposed developing system 2 AI to perform more complex tasks such as reasoning, planning, emotion interpretation, and decision-making. According to the theory of thinking proposed by Daniel Kahneman, there are two systems of thought that influence human cognition: system 1, which is fast and automatic; and system 2, which is slower and more deliberate. System 1 excels at discriminative tasks, while system 2 excels at generative tasks that require more complex reasoning and decision-making. Understanding how the mechanisms and functions of consciousness support thinking and decision-making, we can be well informed to design system 2 AI architecture and components.

To model and develop a system that exhibits human-like consciousness or system 2, this section first reviews the mechanisms of consciousness and survey representative theories and hypotheses proposed by researchers in various fields. It then defines the goals and desired functionality of system-2 AI. While various theories of consciousness have been proposed by researchers in fields of philosophy and theology, we choose to base our modeling efforts on scientific evidence from fields such as physics, biology, neuroscience, and computer science, rather than relying on more abstract and elusive ideas.

A. Definition and Complexity of Consciousness

There have been debates about whether plants and inanimate objects, such as rocks, have consciousness. However, in this paper, we are focused on modeling consciousness in humans. Michio Kaku's definition of consciousness [Kaku(2014)] is simple, understandable, and implementable. According to Kaku, the complexity of an organism's consciousness is determined by the complexity of its sensing and response system. The more complex an organism's ability to sense and respond to stimuli in its environment, the more information is transmitted and processed, leading to a more complex consciousness. Therefore, the complexity of consciousness can be characterized by the complexity of its information processing mechanisms and capacity.

The Integrated Information Theory (IIT) [Tononi(2004)] is similar to Kaku's idea about the relationship between the complexity of an organism's consciousness and its sensory and response system. IIT proposes that consciousness arises from the integration of information across different brain areas, and that the complexity of an organism's consciousness is determined by the amount of integrated information it can process. Other theories of consciousness include the Global Workspace Theory, which suggests that consciousness arises from the interaction between different brain areas, and the Dynamic Core Hypothesis, which proposes that consciousness arises from the interaction of different neural networks in the brain.

Human beings have sensory organs for sight, hearing, smell, taste, touch, and proprioception, which allow us to perceive and interpret stimuli in our environment. This is essential for our survival and our ability to interact with the world.

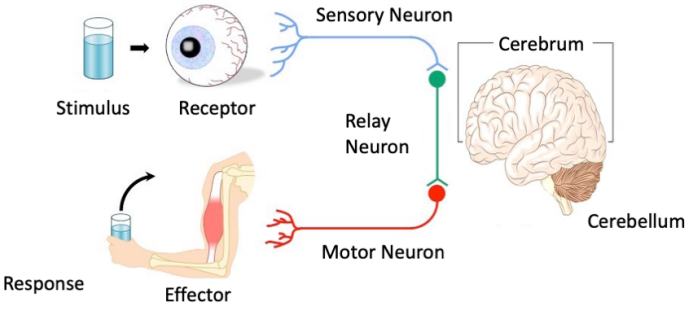

Fig. 1: Bottom-up Attention: Stimulus $\longrightarrow$ Cerebellum $\longrightarrow$ Cerebrum $\longrightarrow$ Response.

How is consciousness aware of changes in our body and the environment? Let's consider the example of a stimulus-response depicted in Figure 1. In this example, the stimulus is a glass of water and the receptor is the human eye. When the eye detects the stimulus, it sends signals to the brain through sensory neurons. The brain processes these signals and makes a plan to fetch the glass of water by issuing movement instructions to the hand (the effector) through motor neurons.

There are two conscious events in this example: the awareness of the sensation of thirst and the process of quenching the thirst. Consciousness is involved in both of these events, but in different ways. The awareness of thirst is an example of bottom-up awareness, which is informed by unconscious processes and rises into consciousness. The process of fetching a glass of water is an example of top-down processing, which involves conscious planning and execution. In the next section, we will explore the mechanisms behind both top-down and bottom-up awareness.

B. Arise of Consciousness

The process of "arising of consciousness" refers to the emergence or appearance of conscious experience, or the process by which certain mental states or events come to be experienced or perceived by an individual. In figure 1, the sensation of thirst is accumulated unconsciously, and when the signal strength reaches a certain threshold, the thirst experience arises into consciousness. This process is facilitated by the unconscious process of accumulating the sensation of thirst.

Sigmund Freud was one of the first to propose a model of the mind that included both conscious and unconscious processes [Freud(1900)]. According to Freud, the unconscious mind is the source of many of our actions and behaviors, and it plays a crucial role in shaping our thoughts and feelings. He believed that the unconscious mind exerts a powerful influence on our conscious thoughts and behaviors.

Since Freud's time, other theories of the unconscious mind have been developed. Carl Jung proposed the concept of the collective unconscious, which he believed was inherited and contained the accumulated experiences and knowledge of our ancestors. He argued that the collective unconscious has a powerful influence on our conscious lives and helps us to better understand the human condition and our place in the world [Jung(1944)].

Unconscious processes also play a crucial role in many vital functions of the human body, such as the regulation of heart rate, respiration, digestion, and other autonomic functions. These processes are often referred to as automatic or reflexive because they occur automatically and do not require conscious thought or awareness. The unconscious mind also plays a role in other aspects of human behavior and cognition, including memory, perception, and even decision-making [Kihlstrom(1987)], [Kihlstrom(1997)], [Peterson(2019)].

While significant progress has been made in understanding the brain and its functions, the precise nature of consciousness and how it arises from the brain and other biological systems is still not fully understood. However, for the purpose of designing and implementing system 2 AI, we believe that the current theories and evidence from fields such as physics, biology, neuroscience, and psychiatry are sufficient for providing guidance to computer scientists. It is worth noting that questions about the nature of consciousness after death and its possible relationship to the soul, while interesting and important in their own right, do not seem to be directly relevant to the current goals of modeling and implementing system 2 AI.

C. Panpsychism vs. Functionalism

There are two theories about consciousness, *panpsychism* and *functionalism*. When considering using these theories to develop a computational model for consciousness, they may not be mutually exclusive.

C.1 Theory of Panpsychism:

Panpsychism is a philosophical theory that posits that consciousness is a fundamental aspect of the universe and present in all matter, including inanimate objects. David Chalmers, Galen Strawson, and Thomas Nagel are some proponents of panpsychism. These philosophers have different viewpoints on the subject, with Chalmers focusing on the problem of explaining the subjective nature of consciousness and its irreducibility, Strawson emphasizing the importance of panpsychism in avoiding the "hard problem" of consciousness[1] and as a framework for understanding the nature of the self and its relationship to the physical world, and Nagel arguing that subjective experience has an ineffable quality that cannot be fully captured by objective descriptions of the physical processes underlying it.

Nagel's views also differ from those of Chalmers and Strawson in two ways: the concept of "what it's like," in which Nagel argues that subjective experience has an ineffable quality that cannot be fully captured by objective descriptions of the physical processes underlying it, and the "bat argument," in which Nagel asserts that there are limits to our understanding of the subjective experiences of other beings, even with a complete scientific understanding of their physical processes. Nagel believes that the subjective nature of experience is a fundamental aspect of the world that cannot be reduced or explained by any physical theory.

While both Nagel and Chalmers recognize the difficulty of explaining subjective experience in physical terms, they have different ideas about how to address this problem. Nagel takes a more skeptical approach, suggesting that subjective experience may be beyond the reach of scientific explanation, while Chalmers is more optimistic and believes that panpsychism may provide a solution.

In addition to these philosophical theories, panpsychism has also inspired more specific models or theories developed by neuroscientists and psychiatrists, including Giulio Tononi's integrated information theory (IIT), Bernard Baars' global workspace theory (GWT), Francis Crick and Christof Koch's neural correlates of consciousness (NCC) approach, and Michael Graziano's attention schema theory (AST). These theories attempt to provide a scientific explanation for the emergence of consciousness and its relationship to the brain and other biological systems. We will discuss relevant details of these models in Section IV when formulating the computational model for consciousness.

C.2 Theory of Functionalism:

Functionalism is a theory of consciousness that proposes that consciousness arises from the function of the brain, rather than its specific physical or neural implementation [Putnam(1967)]. According to this view, consciousness can be understood as a mental or computational process that performs certain cognitive functions, such as perception, attention, decision-making, and so on [Block(1980)].

One key idea of functionalism is that mental states and processes can be described and explained in terms of their causal roles or functions, rather than in terms of their specific neural or physical implementation [Fodor(1968)]. This function-agnostic approach allows a computation model to support the wide variety of different types of conscious experiences that exist, such as the experience of sight, hearing, touching, and so on. Each of these experiences is produced by different neural processes in the brain, but functionalism suggests that they are all instances of consciousness because they all perform similar functions, such as representing the world and guiding behavior [Dennett(1991)]. Therefore, these functions can be supported by the same computational models, such as neural networks [Rumelhart and McClelland(1986)].

A practical benefit for supporting functionalism is that it can account for the fact that consciousness seems to be transferable or multiple realizable [Fodor(1974)]. This is similar to the way a computer program can be run on different types of hardware and still perform the same functions.

Functionalism is often contrasted with panpsychism and other

[1]There is an "explanatory gap" between our scientific knowledge of functional consciousness and its "subjective," phenomenal aspects, referred to as the "hard problem" of consciousness [Koch(2004)].

approaches to the"hard problem" such as substance dualism [Lewis(1966)], [Descartes(1984)], which propose that mental states and processes are inherently subjective and non-physical and cannot be fully explained in terms of physical or neural processes [Nagel(1974)]. However, some argue that the "hard problem" should be directly confronted rather than avoided [Koch(2004)]. For example, different people at different times in different moods may have different feelings and reactions to the same stimulus, such as a snowy mountain or a blue ocean [Solomon and Greenberg(2004)]. These feelings may also depend on personality, memory, and the states of the unconscious mind [Freud(1900)], [Freud(1917)].

D. Empirical Confirmation

We now tie the theoretical models presented in the previous section with the physical brain and the central nervous system (CNS).

Conscious thoughts are processed by the brain. Specifically, the prefrontal cortex, which is the part of the brain responsible for higher cognitive functions such as decision making, problem solving, and planning, is thought to play a key role in the processing of conscious thoughts. The prefrontal cortex is also thought to be involved in the integration of information from various other brain regions, allowing us to make sense of the thoughts and experiences that we have. Other brain regions that are important for the processing of conscious thoughts include the parietal lobe, which is involved in the integration of sensory information, and the temporal lobe, which is important for the processing of language and memory.

One approach to understanding the neural basis of consciousness is to identify the specific brain regions or processes that are necessary for conscious experience. For example, research has suggested that the prefrontal cortex, the thalamus, and the reticular activating system may play a role in conscious processing. However, it is important to note that these brain areas are not the sole source of consciousness, and that many other brain regions and processes likely also contribute to conscious experience.

Other factors that may contribute to consciousness include the activity of neurotransmitters such as dopamine and serotonin, as well as the presence of certain brain waves. Some researchers have also suggested that consciousness may involve the interaction of multiple brain systems, including those involved in perception, attention, and memory.

Another approach to understanding the neural basis of consciousness is to consider how different brain areas and processes work together to support conscious experience. For example, the global workspace theory proposes that conscious experience arises from the integration of information from different brain regions through interactions in the prefrontal cortex.

To verify these theories, Stanislas Dehaene and Jean-Pierre Changeux in [Dehaene and Changeux(2011)] reviewed and discussed a range of experimental studies on the neural basis of conscious processing, including neuroimaging, neurophysiological, and lesion, and transcranial magnetic stimulation studies. More specifically, the experimental studies discussed in the article include:

- Neuroimaging studies using functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) to investigate the brain areas and processes involved in conscious processing [Dehaene and et al.(2001)], [Dehaene and et al.(2006)].
- Neurophysiological studies using electrophysiological techniques, such as electroencephalography (EEG) and magnetoencephalography (MEG), to study the brain activity associated with conscious processing [Dehaene et al.(2003)Dehaene, Sergent, and Changeux], [Dehaene et al.(2011)Dehaene, Jobert, and Naccache].
- Lesion studies, which involve examining the effects of brain damage on conscious processing, to identify the brain areas and processes necessary for conscious experience [Binetti et al.(1998)Binetti, Deecke, Passingham, and Jeannerod], [Dehaene and et al.(1998)].
- Studies using transcranial magnetic stimulation (TMS) to manipulate brain activity and investigate the causal role of specific brain areas in conscious processing [Koch and Tsuchiya(2006)], [Massimini et al.(2005)Massimini, Huber, Ferrarelli, Hill, and Tononi].
- Studies using single-neuron recording techniques in non-human primates to investigate the neural basis of conscious perception and decision-making [Lau and Passingham(2006)], [Quian Quiroga et al.(2005)Quian Quiroga, Reddy, Koch, and Fried].
- Studies using pharmacological manipulations to investigate the role of specific neurotransmitter systems in conscious processing [Dehaene and et al.(2006)].

The experimental studies discussed in the article provide insights into the neural basis of conscious processing and how different brain systems and processes contribute to conscious experience. The authors propose a framework for understanding the neural basis of conscious processing, which involves three levels of analysis: the *neuronal* level, the *network* level, and the *global* level. At the neuronal level, the authors discuss the role of various brain areas, including the prefrontal cortex and the thalamus, in conscious processing. They also discuss the importance of neurotransmitters and oscillations in the brain in supporting conscious processing. At the network level, the authors discuss how different brain areas work together to support conscious processing, and how this involves the integration of information from multiple brain systems. At the global level, the authors discuss how conscious processing depends on the integration of information from multiple sources and how this integration is supported by the prefrontal cortex. They also discuss how this integration allows for the emergence of higher-level cognitive functions, such as attention and decision-making.

Overall, the article provides an overview of how different brain processes and systems contribute to conscious processing, and how this understanding can be used to develop computational models of consciousness.

Section Remarks:
There are two main points to consider when designing a computational model of consciousness.

First, the focus of the model should be on providing the

necessary functions of consciousness to support the desired tasks, such as reasoning, planning, and emotion interpretation and prediction. The specific physical or neural implementation of the model is not as important as its ability to perform these functions. This means that the model does not need to strictly mimic the anatomy and function of the brain in order to be effective.

Second, it is important to address the issue of subjective experience, or the "hard problem" of consciousness, rather than avoiding it. Subjective experience is a critical aspect of consciousness and is essential for many real-world applications, such as the ability to adapt to individuals and their local culture and laws. Ignoring this aspect of consciousness may limit the effectiveness and flexibility of the model. Therefore, it is important to consider how subjective experience can be accounted for and incorporated into the computational model.

III. Functionalities of Human Consciousness

This section list a number of key conscious functions, their specifications, together with concerns about programming them into an artificial agent. In the end of each fuction specification, we remark on its inspirations to the design of our computational model: Computational Consciousness Model (CCM), presented in Section IV.

A. Awareness

Awareness is the ability to be aware of one's surroundings, thoughts, and feelings. In his theory of consciousness, Bernard Baars proposes that consciousness is a global cognitive process that integrates information from various sources and allows an organism to interact with its environment [Baars(1988)]. This process is centered around the concept of a global workspace, which is a hypothetical system in the brain that allows for the integration and availability of information to other cognitive processes. According to Baars, consciousness arises when information is broadcast to the global workspace, making it available for other cognitive processes to act upon.

Baars' theory also includes the concepts of awareness and attention. While awareness and attention are related, they are not the same thing. Awareness encompasses the entire range of conscious experience, while attention is a specific cognitive process that allows an individual to focus on certain stimuli or sources of information.

Researchers have attempted to model the global workspace in several ways, including through computer simulations and brain imaging studies. One example of a simulation model of the global workspace is the *blackboard* architecture, which was designed to replicate the cognitive processes involved in the global workspace, including the integration of information from various sources and the ability to broadcast this information to other cognitive processes. Other researchers have used brain imaging techniques, such as functional magnetic resonance imaging (fMRI)[2], to study the neural basis of the global workspace and to test predictions of Baars' theory. For example, some studies have found that certain brain regions, such as the prefrontal cortex and the posterior parietal cortex, are more active when an individual is engaged in tasks that involve conscious processing, which is consistent with Baars' theory.

[2]The brain probing and visualization techniques have been enhanced dramatically since the invention optogenetics by K. Deisseroth in 2000 [Deisseroth(2021)]

Notes to CCM design

In CCM, an event that is being aware of can be placed in a task/job pool, awaits for a central scheduler to prioritize and pay attention to. We discuss the *attention* function and its mechanisms next.

B. Attention, Bottom-Up and Top-Down

Attention is the ability to focus on specific stimuli or tasks, and to filter out distractions [Baars(1988)]. Attention allows us to efficiently process and attend to important information and tasks, and to ignore irrelevant or distracting stimuli. Attention is also closely linked to our perception, memory, and decision-making processes, as the information that we attend to is more likely to be encoded in memory and to influence our decisions.

Posner and Petersen propose a model of the attention system in [Posner and Petersen(1990)] based on evidence from various sources, including behavioral studies, brain imaging studies, and studies of brain-damaged patients. Their model consists of three interacting components: the alerting system, the orienting system, and the executive system. The alerting system maintains an overall state of alertness and arousal, while the orienting system directs attention to specific stimuli in the environment. The executive system controls the allocation of attention and coordinates the activity of the other two systems.

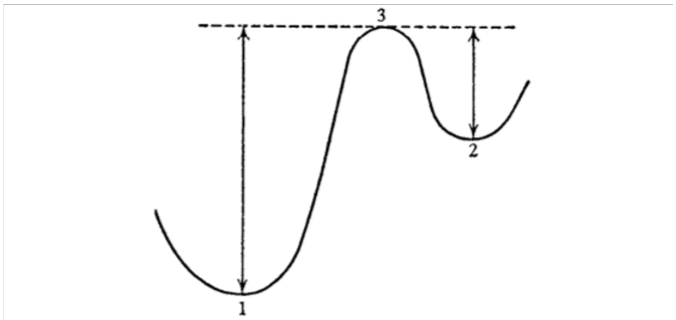

Fig. 12. Energy threshold (3) between the isomeric levels (1) and (2). The arrows indicate the minimum energies required for transition.

Fig. 2: A jump may occur when energy peaks at point 3.

Erwin Schrödinger's attention model, described in his book "What is Life" [Schrödinger(1944)], suggests that the attention mechanism works in a way similar to a "quantum jump" in quantum mechanics. (Figure 2 illustrates a quantum jump as described in Schrödinger's book.) According to Schrödinger's model, information is constantly being received by the sense organs and processed by the unconscious mind. When the accumulated energy of certain signals (e.g., heat) reaches a threshold, a quantum jump is triggered, and the conscious mind becomes aware of the new event. The conscious mind prioritizes its attention by examining the pending alerts in its

executive system and scheduling the top priority task to be handled by the orient system.

Once a person is already in attention mode, they can plan their next action and inform the relevant effectors (such as their arms, legs, or sense organs) to act or further sense. This is the top-down attention process taking place entirely in the conscious mind.

Schrödinger's model is attractive because it connects to physics and can be implemented using current computing functions. It also explains the transition from consciousness to unconsciousness through the second law of thermodynamics [Schrödinger(1944)]. The model suggests that the threshold for triggering a quantum jump in consciousness can be influenced by various factors, such as the intensity of the sensory input and the overall energy level of the system. This aligns with the idea that attention is a limited resource that can be influenced by various factors, such as motivation, fatigue, and prior experience. In this way, Schrödinger's model provides a possible physical basis for the attention mechanism and the dynamic nature of consciousness.

Notes to CCM design This mental model of attention and the bottom-up unconsciousness to consciousness) and top-down consciousness to consciousness) mechanisms can be formulated as a task/job scheduling algorithm in the operating system of an AI agent. We discuss design considerations such as priority determination and quantum setting of tasks in Section IV.

C. Perception

Perception is the process of interpreting sensory information and forming mental representations of the environment [Gregory(1997)]. This process is typically supported by system 1 AI, or unconsciousness. However, a computational model should consider how the transitions between unconscious background perception and consciousness awareness are performed. Schrödinger's work provides insights into the mechanisms in physics that could be used to implement these transitions, as described under the attention function.

D. Thinking

Thinking is the ability to process and manipulate information, solve problems, and make decisions. In psychology, thinking refers to the mental process of generating, manipulating, and evaluating ideas and concepts. It involves actively constructing, organizing, and evaluating information, as well as creating new ideas and problem-solving.

In their book "Human Problem Solving," Allan Newell and Herbert Simon propose a theoretical framework for understanding the process of thinking and problem-solving [Newell and Simon(1972)]. They argue that problem-solving involves searching for and generating new knowledge, and that this process can be broken down into several distinct stages: formulation, search, evaluation, and execution.

- Formulation: This stage involves defining the problem and understanding its context and constraints.
- Search: During this stage, the problem-solver generates and considers potential solutions to the problem.
- Evaluation: In this stage, the problem-solver evaluates the potential solutions and selects the one that is most likely to be successful.
- Execution: This stage involves implementing the chosen solution and verifying that it has solved the problem.

There are various theories and models in psychology that attempt to explain the process of thinking and how it can be influenced by different factors. Some models relevant to our design purpose are: the dual process model [Kahneman(2011)], the information processing model [Miller(1956)], the cognitive psychology model [Newell and Simon(1972)], the connectionist model [Rumelhart and McClelland(1986)], and the social cognitive theory [Bandura(1977)].

The dual process model proposes that there are two distinct systems of thinking: system 1, which is intuitive and automatic, and system 2, which is more deliberative and controlled. The AI community has recognized the system 1 vs system 2 characterization, and so does this paper. The information processing model proposes that thinking involves manipulating and transforming information through stages such as perception, attention, and memory. The transformer architecture and large pre-trained language models (LLMs) well represent these stages. (We will present their merits shortly.) The cognitive psychology model emphasizes the role of mental representations and processes in thinking and problem-solving, while the connectionist model proposes that thinking and learning occur through the formation and strengthening of connections between neurons in the brain. The social cognitive theory emphasizes the role of social and cultural factors in shaping thinking and behavior. Some core deep learning algorithms, such as convolutional neural networks (CNNs), learn mental representations of information from training data. Since the training data can adapt to social and cultural factors, the social cognitive theory can be incorporated through data. The weights between neurons in connectionist models can characterize the strength of connections between neurons and can be flexibly changed to adapt to environmental changes [Russell and Norvig(2010)], [Laudon and Laudon(2016)].

The transformer architecture, introduced in [Vaswani et al.(2017)], allows for the meaning of an entity, such as a word in a document or a blob in an image, to be determined by both the entity itself and its context, which is represented by neighboring entities. This context-based semantic resolution method has significantly improved the performance of natural language processing (NLP) applications. In addition, large pre-trained models (e.g., [Brown(2020)]) with a high number of layers and neurons can serve as a comprehensive knowledge base for fine-tuning and prompting when used for tasks such as language translation [Devlin et al.(2018)Devlin, Chang, Lee, and Toutanova] and ChatGPT [OpenAI(2021)].

Recently, the use of fine-tuning methods and prompting mechanisms for large pre-trained language models (LLMs) has gained significant attention in the machine learning community (e.g., [Gao(2021)]). One such approach is the chain of thought method [Wei et al.(2022)Wei, Wang, Schuurmans, Bosma, Chi, Le, and Zhou], which aims to mimic the human thinking process to prompt LLMs and improve their ability to perform complex reasoning tasks. Another example is our own work, in

Fig. 3: Free Will? Adam and Eve, Rembrandt (1606-69).

which we use Socratic dialogue[3] to enhance the effectiveness of prompting with LLMs [Chang(2022b)]. These efforts highlight the ongoing efforts to improve the performance of LLMs in a variety of tasks.

Notes to CCM design

A thinking process or a problem-solving session requires a knowledge base, which can be served by large pre-trained language models (LLMs). The four-step thinking process can be modeled as follows:

- Formulation: The question or problem is formulated into an input prompt template, which specifies the objective(s), criteria, and individual, social, and cultural preferences.
- Search: A search is conducted on LLMs, and the results are cross-validated and aggregated.
- Evaluation: The potential solutions generated from the search are evaluated against the criteria specified in the input prompt template.
- Execution: The chosen solution is implemented and verified to see if it successfully solves the problem.

E. Free Will and Guardrails

The ability to choose goals and to act to achieve those goals is free will [Dennett(1987)]. The debate between free will and determinism is a longstanding and complex one, with strong arguments on both sides. (Figure 3 presents Rembrandt's masterpiece "Adam and Eve" symbolizing this struggle.) While determinism holds that all events, including human actions and behaviors, are predetermined by external factors such as genetics, identity, and environment (e.g., [Dennett(2003)]), proponents of free will argue that individuals have the ability to make choices and act on them freely (e.g., [Kane(1996)]), such as choosing our words and actions, deciding how to spend our free time, and more seriously, selecting a candidate to vote for. While it is true that certain aspects of an individual's life, such as our DNA and identity, may not be within our control, it seems that there is still room for free will in the choices and actions that individuals make on a day-to-day basis.

In the context of AI, the question of whether an AI agent could have free will depends on how the concept of free will is understood and defined. If free will is understood as the ability to make choices and act on them freely, without being predetermined by external factors, it is unlikely that an AI agent could have free will in the same way that a human does. This is because AI systems are typically designed to follow specific rules or algorithms in making decisions, and their actions are not independent of these rules or the data and input that they are provided with (e.g., [Russell and Norvig(2010)]). However, it is possible for an AI system to be designed to make decisions in a more flexible or adaptive way, especially when the cost of making a bad choice is low. For example, a robot may be able to answer a typical question during dating, such as "Who would you save first in a fire, me or your parents?" In this case, the AI agent could be said to have a certain degree of autonomy, in the sense that it is able to make decisions and take actions based on its own evaluation of the situation, rather than simply following predetermined rules.

In situations where the cost of making a bad decision is high, such as a self-driving vehicle faced with a potential crash, the question of free will becomes more complex. In these cases, the decision made by the AI agent would still be based on the specific criteria and rules it has been programmed with, rather than on any sense of free will. When the cost of a bad choice is not insignificant, free will may be permitted. The higher the stakes of a decision, the lower the risk that an AI agent can take. Additionally, an AI agent must be able to adapt to its environment and consider external rewards and penalties when making a decision. Human values and preferences, such as not tolerating inappropriate language, must also be considered in the decision-making process.

Regarding implementation, free will can be formulated by the rewards and costs of choices and the entropy among choices [Rehn(2022)]. This means that an individual's free will is represented by the consequences of various options they have to choose from and the inherent uncertainty in their decision-making process.

F. Emotion, Expression and Interpretation

Emotion is an experience of feelings that can take place unconsciously or/and consciously. A sudden outburst of emotion is largely irrational and is reflective without passing through consciousness evaluation.

[3]Socratic dialogue is a method of questioning and discussion that was developed by the ancient Greek philosopher Socrates. It is a form of inquiry in which one person (the interlocutor) asks questions of another person (the respondent) in order to clarify their thoughts and beliefs, and to help them arrive at a deeper understanding of a topic. In Socratic dialogue, the interlocutor typically plays the role of the teacher, while the respondent plays the role of the student. The goal of the dialogue is not necessarily to arrive at a definitive answer or solution, but rather to explore different perspectives and ideas in order to gain a deeper understanding of a subject. Socratic dialogue has been used as a teaching method for centuries and is still widely used today in education and other contexts.

While the idea of programming emotions into artificial agents may be a controversial topic, there are certainly benefits to be gained from the ability to convey care, understanding, and support through facial expressions and other forms of nonverbal communication. Antonio Damasio's work in "Descartes' Error" [Damasio(1994)] highlights the importance of emotions in guiding human decision-making and influencing our sense of self and perception of the world. It is not unreasonable to believe that these same emotions could be useful for artificial agents in building meaningful and effective relationships with humans.

In fact, a survey conducted at a senior home for a case study on caring behaviors found that certain emotions and expressions were particularly comforting and desirable to seniors. These included being attentive, love, empathy, joy, and laughter, as well as expressions of gratitude and appreciation that brought a sense of contentment and happiness.

- Being attentive: calling one's name and remembering previous conversations.
- Love: feeling loved and cared for by others can be a great source of comfort and support.
- Empathy: feeling understood and supported by others can be very comforting, especially when one is going through a difficult time.
- Joy: experiencing joy and happiness can bring a sense of well-being and comfort.
- Humor: sharing a good laugh with others can help to lighten the mood and provide a sense of relief and relaxation.

Expressing gratitude and appreciation can also bring a sense of contentment and happiness. In terms of facial expressions, a warm and friendly smile is often comforting, as is a look of concern or understanding. Overall, the most comforting emotions and expressions are those that convey care, understanding, and support.

The use of large pre-trained language models and prompting mechanisms can enable the programming of emotions in artificial agents. Facial expressions, such as a warm and friendly smile or a look of concern or understanding, can also be effective in conveying comfort and support. It may be beneficial for an artificial agent to be able to adjust its demeanour based on the reactions of individuals it interacts with.

Section Remarks:

These functionalities are not necessarily distinct from one another, and they often overlap and interact in complex ways. Additionally, there may be other functionalities or characteristics of consciousness that are not captured by this list. Understanding the functionalities of human consciousness is a complex and ongoing scientific challenge.

While it is possible to program robots to support this list of consciousness behaviors, it is important to consider the potential limitations and ethical implications of doing so. For example, some may argue that it is not appropriate to try to replicate human emotions in a machine, or that it could create false expectations or misunderstandings if the robot's emotional responses are not genuine. It may also be necessary to consider the potential consequences of programming robots with certain emotions or behaviors, such as the potential for misuse or abuse. Regarding free will, some may consider giving any freedom to any artificial agents should be prohibited because of both ethical and legal concerns.

Ultimately, the decision to program robots with conscious behaviors will depend on the specific goals and context of the application, as well as the values and ethical considerations of those involved.

IV. Computational Consciousness Model (CCM)

The Computational Consciousness Model (CCM) consists of four modules: the receptor, unconsciousness, consciousness, and effector modules.

We present the Computational Consciousness Model (CCM), which consists of four logical modules: the receptor, unconsciousness, consciousness, and effector modules. These modules are illustrated in the stimulus-response diagram in Figure 1.

- The receptor module processes input signals from sensors, such as sight and sound, and converts them into representations that are sent to the global workspace of the unconsciousness module.
- The unconsciousness module performs discriminative classification on the received representations and acts as a scheduler, maintaining a list of pending events and their energy levels using a multi-level feedback queue.
- The consciousness module is single threaded and executes one process or task at a time, maintaining a schema as suggested by the AST model. The schema maintains the states of each task, which depend on the source receptor. For example, the state of seeing that is currently being processed in the consciousness module can receive a top-down attention signal to orient the sensory processing and zoom in on the stimuli. This signal, along with a set of new parameters, is then sent to the corresponding effectors (e.g., the eyes).
- The effectors are reactive and wait for signals from the consciousness module to act according to the provided parameters. An effector can also act as a receptor, sending feedback signals to the unconsciousness module.

The consciousness module is the only component that requires further investigation. It consists of three sub-components: a scheduler and its related data structures, an external reward system, and an intrinsic reward system. The consciousness module not only maintains its own state, but also the states of aware stimuli in the environment, such as people and objects.

A. Scheduler

The Computational Consciousness Model (CCM) uses a multi-level feedback queue (MFQ) scheduler [Wikipedia(2021)] to manage tasks. In MFQ, processes and tasks are organized into a hierarchy of queues, with each queue having a different priority level. The scheduler selects the task at the front of the highest priority queue that has pending tasks in it. If a task doesn't finish within its time slice (also known as the quantum size), it is moved to the back of the queue at its current priority level. If a process finishes, it is removed from the queue.

Traditional MFQs in operating systems are configured with different quantum sizes for each priority level, with shorter

sizes for higher priority queues and longer sizes for lower priority queues. This allows tasks in higher priority queues to be serviced sooner and more frequently, while giving tasks in lower priority queues the opportunity to run if the higher priority queues are empty. When a task that is not in the highest priority ends its quantum, it is promoted to the next higher priority queue. The MFQ scheduler adjusts four parameters to meet two requirements: fairness, where tasks are executed based on their importance; and starvation-free, where every task will eventually be executed. The four tunable parameters of MFQ are: 1) the number of priority levels, 2) the quantum size of each level, 3) task promotion policy, and 4) task demotion policy.

Let us examine and answer four design questions for CCM-MFQ: First, what are the conscious capabilities that the MFQ scheduler must support? Second, how these parameters ought to be set to manage tasks in conscious and unconsciousness states? Third, in addition to fairness and starvation-free, are there additional policies to be added to MFQ?

Conscious Capabilities

Section III specifies six functionalities including awareness, attention, perception, thinking, free will with ethics, and emotion expression and prediction. Of these, awareness and attention are the states of a task, which can be directly supported by the multi-level feedback queue (MFQ) system.

The CCM's MFQ scheduler can be thought of as the brain of the global working space, managing all tasks in the unconscious and conscious states. Tasks in the unconscious state are all parked in the lowest priority level queue in any order. Tasks in all levels of the MFQ except for the bottom, lowest priority level are considered *aware tasks*. The current task being executed is the *attended* task. The functionalities of awareness and attention can be directly supported by the MFQ.

The specifications for the other four functionalities are represented by computer-executable tasks that are scheduled in the conscious-level queues. Their priorities are determined by their reward values, both external and intrinsic.

Parameter Settings

To determine the parameters for the MFQ scheduler in the CCM to manage tasks in the conscious and unconscious states, it is important to consider the relative importance and time sensitivity of these tasks. For example, tasks related to maintaining an individual's safety or well-being, such as controlling vital body functions, should be given higher priority and shorter time slices. Tasks related to long-term goals or planning can be given lower priority and longer time slices, as they can be completed over a longer period of time without significant impact.

Additional Mechanisms

The CCM must implement an interrupt mechanism to support the transition from unconsciousness to consciousness. When a task in the unconscious state reaches the energy threshold, it interrupts the scheduler to place it in one of the high-level queues based on its priority.

To support inter-task synchronization, additional policies may be necessary to ensure that tasks are completed in a specific order or dependent on the completion of other tasks. For example, in an eye-hand coordination task involving multiple receptors and effectors, the master task may synchronize with sight receptor and hand effector tasks to execute simultaneously or in a predetermined order.

It is important to note that the global workspace may need to support asynchronous tasks executing on a distributed, multi-process computing environment.

B. External Reward System for Adaptivity

Using rewards to train an AI agent to behave optimally and achieve a maximum total reward is a common approach in reinforcement learning [Sutton and Barto(2018)]. This approach can be effective at shaping the behavior of an AI agent in a desired way and helping it to adapt to different circumstances. For example, if an AI agent is designed to be a caregiver at a senior home, the priorities of routine tasks can be determined by the supervisors. Once task rewards are set, they can be scheduled to their corresponding priority queues in the MFQ. In our prior work on healthcare disease diagnosis [Peng et al.(2018)Peng, Tang, Lin, and Chang], we used reinforcement learning and reward/feature shaping to adapt to user feedback. This computational framework allows us to refine reward values and reshape the feature space in order to better meet the needs and preferences of individuals.

However, rewards related to emotion cannot be handled by the MFQ alone, as this requires input from the user. For example, if we want our caregiver AI to be empathetic, the specification of empathy must come from a list of instructions provided by the user. When a behavior is rewarded or complained about by the user, that behavior can be reinforced or discouraged. Humor is another example where adaptation must come from user feedback.

An AI agent can become more adaptive to its users and environment through learning from demonstrations, where the agent is trained to imitate the actions of a human expert or teacher. This method can be effective at transferring knowledge and skills from humans to the AI agent, especially when it is difficult to explicitly specify the desired behavior of the agent in terms of a reward function. The use of large pre-trained language models (LLMs) allows for demonstrations to be provided via prompts, which can serve as templates that include descriptions of goals and specific, focused instructions with examples.

In summer 2022 at Stanford, we initiated the *Noora* chatbot project, which aims to help autism patients to learn how to speak with empathy by providing templates that include instructions for comforting and harmful responses. A sample template starts with an instruction like this:

"Hi Noora, I'm reaching out to you because you are a good friend and I value your support and understanding. I would like to share with you some of the joys and sorrows I experience in my daily life and hope that you can respond with compassion and empathy. Below, I've provided some example dialogues

to illustrate what I consider to be comforting and harmful responses. Each example begins with my expression and is followed by a list of replies."

The first example dialogue starts with a statement: *"I was laid off by my company today!"* followed by a sample list of good and a list of bad responses.
Empathetic responses:

- "I'm so sorry to hear that. Losing your job can be a really tough and stressful experience. How are you doing?"
- "That must have been a really difficult and unexpected news. I'm here to listen and support you however I can."
- "I can imagine how hard and unsettling it must have been to receive that news. Is there anything you'd like to talk about or anything I can do to help?"

Non-empathetic responses:

- "That's too bad, but there are plenty of other jobs out there. You'll find something soon enough."
- "Well, you probably weren't very good at your job if they let you go."
- "I don't know why you're so upset about this. It's not like it's the end of the world."

In order to improve the adaptability of the Noora chatbot, it is important for the trainer or supervisor to provide as many examples as possible for the agent to learn from. This will allow the Noora chatbot to better understand how to respond with empathy to novel statements. Additionally, it is important for the user to provide feedback on the chatbot's responses, particularly when they are not empathetic. This can help the chatbot to understand what kind of responses are expected and desired, such as those that acknowledge the user's feelings and offer support and understanding. The user can also reward the chatbot with positive facial expressions or laughter in response to jokes to encourage empathetic behavior. Overall, this combination of training and feedback can help the Noora chatbot to become more adaptive and empathetic in its interactions with users.

C. Intrinsic Reward System for Ethics

In order to shape the behavior of an AI agent in an ethical manner, intrinsic rewards can be used in reinforcement learning. These rewards serve as positive reinforcement for actions that align with ethical values and principles, and allow the agent to derive satisfaction and fulfillment from making its users happy and fulfilled.

To teach desired behaviors and ethics to an AI agent through demonstrations, the intrinsic reward system can utilize prompting templates similar to those used for external rewards. For example, the template for empathy can be used to model other positive behaviors such as being attentive and caring (see the list of positive emotions in Section III). It's important to note that a machine may already possess some positive characteristics (e.g., infinite amount of patience), but we still need to explicitly model good and bad behaviors for the agent to understand and effectively interact with human users. Negative behaviors may include being unpleasant, rude, greedy, lazy, jealous, or prideful, and engaging in sinful or deceitful actions. By using prompting templates and soliciting user feedback, the intrinsic reward system can be adapted to the individual and their local culture and laws.

It is important to note that it is not only the AI agent that should be trained to follow ethical codes, but the supervisors and users should also be held to these standards. The agent should be able to perceive and evaluate the behavior of these individuals to ensure that they are acting in an ethical manner.

D. Developing Prompting Templates

Socratic dialogue is a method of inquiry and critical thinking developed by the ancient Greek philosopher Socrates. It involves asking questions and engaging in dialogue with others in order to explore and clarify ideas, expose contradictions, and arrive at a deeper understanding of a topic. The Socratic method has been influential in the field of philosophy and has also been applied in education and other fields as a way of fostering critical thinking and intellectual curiosity.

There have been many works that have studied or discussed the Socratic method and its influence on philosophy and education. A couple good references for our future research include:

- "The Socratic Method" by Ward Farnsworth, which is a dialogue that discusses the Socratic method and its role in teaching [Farnsworth(2021)].
- "Circles of Learning: Applying Socratic Pedagogy to Learn Modern Leadership" by Katherine L. and Clinton M. Stephens, which discusses the use of Socratic dialogue as a pedagogy for transformative learning and suggests that it can be an effective way to promote critical thinking and encourage students to take an active role in their own learning [Friesen and Stephens(2019)].

V. Case Study: Design a Long-term Care Robot

There are several key challenges facing senior care and long-term care (nursing home care) in the United States and abroad, including issues of affordability[4], accessibility, quality of care, and even instances of abuse and neglect. Care-giving can also be a physically and emotionally demanding task, and caregivers may not have sufficient support or resources to manage the burden.

In order to address these challenges and improve the quality of care for seniors, we plan to develop an artificial agent that could assist with various day-to-day caring tasks. The design process follows these steps:

- Identifying key requirements for a healthcare artificial agent.
- Conducting a feasibility assessment on the top five requirements on the list.
- Identifying three requirements that could potentially be addressed by existing system-1 AI capabilities, and explaining how these could be implemented.
- Selecting one "hard" requirement to tackle through the development of system-2 AI capabilities.
- Sketching out our design and technical tasks, and working towards implementation.

[4](Figure 4 shows the monthly cost by state in the US in 2022.

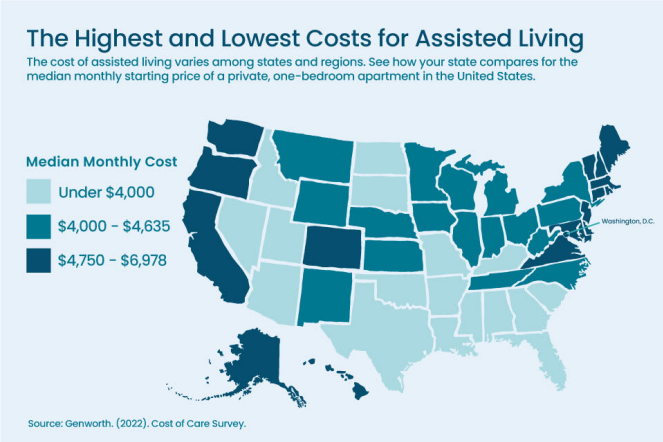

Fig. 4: US Assistant Living Cost by State, 2022. (Source: Genworth. (2022). Cost of Care Survey.)

- Making the design adaptive to personal needs and preferences.

At the top level, we would like an AI agent to help a care recipient become a better version of themselves day by day. This can be achieved through improving various aspects of the care recipient's life, such as their health, happiness, and sense of beauty. These goals can be communicated to the AI agent through prompting templates that provide positive and negative examples. Improving the effectiveness of these templates is a critical research area for this case study.

To meet these goals, the AI agent should have certain functional capabilities, including the ability to perform preliminary diagnosis on the care recipient's health condition, access a knowledge base of relevant diseases and conditions, and display desired attributes such as attentiveness, empathy, patience, love, and humor. Any health concerns identified by the AI agent should be escalated to human supervisors through defined protocols.

To fulfil this list of requirements, we consider that using SCM with medical IoTs as extended signal receptors is sufficient to form a hardware baseline. Regarding medical domain knowledge and disease diagnosis and prediction, various vital-signing monitoring, diagnostic algorithms, and remedy recommendations have been developed and are ready to interoperate. Please consult an extended survey including our own healthcare documented in [Chang(2022a)].

VI. Concluding Remarks

The aim of this paper is to develop a comprehensive model of consciousness for system-2 AI, which can perform tasks involving reasoning, planning, and decision making. To do this, we have reviewed the principles established by philosophers and the theories developed by psychiatrists and neuroscientists based on observations from various empirical studies. While our understanding of consciousness, including its nature, location in the brain, and the way different brain regions interact through the central nervous system, is still incomplete, we are encouraged by the idea of functionalism, which suggests that the implementation of consciousness in machines does not have to follow human anatomy.

Based on the widely accepted theories and principles, including GWS, ITT, DCH, and AST, we have proposed the Computational Consciousness Model (CCM), which is composed of four modules: receptor, effector, unconsciousness, and consciousness, and three subsystems: a scheduler, an external reward system, and an internal reward system. We have demonstrated how these subcomponents can be formulated using well-tested scheduling and reinforcement learning algorithms. To determine and calibrate reward values for both external and internal rewards, we have suggested using prompting templates and soliciting user feedback to make the reward systems adaptive to the individual and his or her local culture and laws, in order to support subjectivity.

There is still much work to be done to ensure that computational consciousness can be effectively and safely deployed. Some key areas of research include:

- Examining the issue of free will and how it can be ethically and safely modeled within an AI system.
- Developing AI agents that not only understand and predict their own states, but also the states of their users and the surrounding environment in order to effectively interact.
- Improving the prompting mechanism to be more effective in shaping the behavior of AI agents.
- Gaining a deeper understanding of how the human brain and nervous system work together to support conscious experience. Techniques such as optogenetics may provide new insights that can be applied to the development of computational consciousness. (See Appendix A for an introduction to optogenetics.)

Acknowledgment

This article was written based on the author's lecture notes of Stanford CS372 from 2020 to 2022 [Chang(2020-22)].

The author would like to thank the ChatGPT Assistant for providing helpful feedback and suggestions during the writing process. (This statement was provided by ChatGPT.) ChatGPT provides the following specific assistance:

- Helping editing paragraphs.
- Recommending a useful reference: the AST model, which suggests that an AI agent should keep track of not only its self-state but also the users' state.

References

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

[Bandura(1977)] Albert Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191, 1977.

[Bengio(2020)] Yoshua Bengio. The future of ai: Opportunities and challenges. *Nature*, 579(7798):479–482, 2020.

[Binetti et al.(1998)Binetti, Deecke, Passingham, and Jeannerod] G Binetti, L Deecke, RE Passingham, and M Jeannerod. Cortical control of saccades. *Experimental Brain Research*, 121(1):66–74, 1998.

[Block(1980)] Ned Block. What is functionalism? *The Journal of Philosophy*, 77(2):5–22, 1980.

[Brown(2020)] Tom B. et al. Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[Chang(2020-22)] Edward Y Chang. Consciousness modeling lecture series. CS372, Stanford University, 2020-22, 2020-22. URL http://infolab.stanford.edu/~echang/cs372/cs372-syllabus.html.

[Chang(2022a)] Edward Y. Chang. Knowledge-guided data-centric ai in healthcare: Progress, shortcomings, and future directions, 2022a. URL https://arxiv.org/abs/2212.13591.

[Chang(2022b)] Edward Y. Chang. Towards artificial general intelligence via consciousness modeling (invited talk). In *IEEE Infrastructure Conference*, September 2022b. URL https://drive.google.com/file/d/1NPuKPB4gSeJeT1fmfY5eus_Rw3abwd5m/view?usp=sharing.

[Crick and Koch(2003)] Francis Crick and Christof Koch. The neural correlates of consciousness. *Nature Neuroscience*, 6(2):119–126, 2003.

[Damasio(1994)] Antonio R Damasio. *Descartes' error: Emotion, reason, and the human brain*. New York, NY: Putnam, 1994.

[Dehaene and Changeux(2011)] Stanislas Dehaene and Jean-Pierre Changeux. Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227, 2011.

[Dehaene and et al.(1998)] Stanislas Dehaene and et al. Cerebral pathways for word masking and unconscious repetition priming. *Nature neuroscience*, 1(7):620–625, 1998.

[Dehaene and et al.(2001)] Stanislas Dehaene and et al. Imaging unconscious semantic priming. *Nature*, 415(6869):26–27, 2001.

[Dehaene and et al.(2006)] Stanislas Dehaene and et al. Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, 10(5):204–211, 2006.

[Dehaene et al.(2003)Dehaene, Sergent, and Changeux] Stanislas Dehaene, C Sergent, and Jean-Pierre Changeux. A neural network model of the basal ganglia's role in saccade initiation. *Nature Neuroscience*, 6(5): 450–459, 2003.

[Dehaene et al.(2011)Dehaene, Jobert, and Naccache] Stanislas Dehaene, A Jobert, and L Naccache. Experience-dependent neural integration of letter strings in the ventral visual pathway. *Nat Neurosci*, 14(9): 1449–1455, 2011.

[Deisseroth(2021)] Karl Deisseroth. *Projections: The Future of the Brain*. Penguin Press, 2021.

[Dennett(1987)] Daniel C Dennett. *The intentional stance*. MIT Press, Cambridge, MA, 1987.

[Dennett(1991)] Daniel C Dennett. *Consciousness explained*. Little, Brown and Company, 1991.

[Dennett(2003)] Daniel C Dennett. Freedom evolves. *Penguin*, 2003.

[Descartes(1984)] Ren'e Descartes. *Meditations on first philosophy*. Hackett Publishing, 1984.

[Devlin et al.(2018)Devlin, Chang, Lee, and Toutanova] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Farnsworth(2021)] Ward Farnsworth. *The Socratic Method: A Practitioner's Handbook*. Godine, Boston, 1 edition, October 2021.

[Fodor(1974)] Jerry Fodor. Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, 28(2):97–115, 1974.

[Fodor(1968)] Jerry A Fodor. Psychological explanation: An introduction to the philosophy of psychology. *Random House*, 1968.

[Freud(1900)] Sigmund Freud. *The interpretation of dreams*. Macmillan, New York, 1900.

[Freud(1917)] Sigmund Freud. *Introductory lectures on psycho-analysis*. Norton, New York, 1917.

[Friesen and Stephens(2019)] Katherine L Friesen and Clinton M Stephens. *Circles of Learning: Applying Socratic Pedagogy to Learn Modern Leadership*. Iowa State University, 2019.

[Gao(2021)] Tianyu Gao. Prompting: Better ways of using language models for nlp tasks. *The Gradient*, 2021.

[Graziano(2016)] Michael S Graziano. Attention schema theory: A mechanistic theory of subjective awareness. *Trends in cognitive sciences*, 20 (8):588–600, 2016.

[Graziano(2013)] Michael S A Graziano. *Consciousness and the social brain*. Oxford University Press, 2013.

[Gregory(1997)] Richard L Gregory. *Eye and brain: The psychology of seeing*. New York, NY: Oxford University Press, 5 edition, 1997.

[Jung(1944)] C. Jung. *Psychology and Alchemy*. Princeton University Press, 1944.

[Kahneman(2011)] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[Kaku(2014)] Michio Kaku. *The future of the mind: The scientific quest to understand, enhance, and empower the mind*. Doubleday, 2014.

[Kane(1996)] Robert Kane. The significance of free will. *Oxford University Press*, 1996.

[Kihlstrom(1987)] John F Kihlstrom. The cognitive unconscious. *Science*, 237(4821):1445–1452, 1987.

[Kihlstrom(1997)] John F Kihlstrom. The cognitive unconscious. In *The new unconscious*, pages 43–65. Oxford University Press, 1997.

[Koch(2004)] Christof Koch. The "hard problem" of consciousness. *Nature*, 467(7319):1121–1122, 2004.

[Koch and Tsuchiya(2006)] Christof Koch and Naotsugu Tsuchiya. Magnetic resonance imaging of the conscious human brain. *Philosophy, Ethics, and Humanities in Medicine*, 1(1):4, 2006.

[Koch and Tsuchiya(2012)] Christof Koch and Naotsugu Tsuchiya. Neural correlates of consciousness: An update. *Annual Review of Neuroscience*, 35:79–97, 2012.

[Lau and Passingham(2006)] H Lau and RE Passingham. Dissociable roles of lateral and medial orbitofrontal cortex in decision-making. *Cortex*, 42(4):393–405, 2006.

[Laudon and Laudon(2016)] Kenneth C Laudon and Jane P Laudon. *Management information systems: Managing the digital firm*. Pearson Education, Upper Saddle River, NJ, 15 edition, 2016.

[Lewis(1966)] David Lewis. An argument for the identity theory. *The Journal of Philosophy*, 63(1):17–25, 1966.

[Massimini et al.(2005)Massimini, Huber, Ferrarelli, Hill, and Tononi] Marcello Massimini, Reto Huber, Fabio Ferrarelli, Sean Hill, and Giulio Tononi. Breakdown of corticocortical connections during sleep. *Science*, 309(5744):2228–2232, 2005.

[Miller(1956)] George A Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81–97, 1956.

[Nagel(1974)] Thomas Nagel. What is it like to be a bat? *The Philosophical Review*, 83(4):435–450, 1974.

[Newell and Simon(1972)] Allen Newell and Herbert A Simon. *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall, 1972.

[OpenAI(2021)] OpenAI. Chatgpt, 2021. URL https://openai.com/blog/chatgpt/.

[Peng et al.(2018)Peng, Tang, Lin, and Chang] Yu-Shao Peng, Kai-Fu Tang, Hsuan-Tien Lin, and Edward Chang. REFUEL: Exploring sparse features in deep reinforcement learning for fast disease diagnosis. In *Advances in Neural Information Processing Systems*, pages 7333–7342, 2018.

[Peterson(2019)] J. Peterson. *Beyond Order: 12 More Rules for Life*. Random House, 2019.

[Posner and Petersen(1990)] Michael I Posner and Steven E Petersen. The attention system of the human brain. *Annual Review of Neuroscience*, 13:25–42, 1990.

[Putnam(1967)] Hilary Putnam. Psychological predicates. *Art, Mind, and Religion*, pages 37–48, 1967.

[Quian Quiroga et al.(2005)Quian Quiroga, Reddy, Koch, and Fried] Rodrigo Quian Quiroga, L Reddy, C Koch, and I Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045): 1102–1107, 2005.

[Rehn(2022)] Emil M Rehn. Free will belief as a consequence of model-based reinforcement learning. *arXiv:2111.08435v2*, 2022.

[Rumelhart and McClelland(1986)] David E Rumelhart and James L McClelland. Parallel distributed processing. *Parallel distributed processing*, 1: 45–76, 1986.

[Russell and Norvig(2010)] Stuart Russell and Peter Norvig. *Artificial intelligence: A modern approach*. Prentice Hall, Upper Saddle River, NJ, 3 edition, 2010.

[Schrödinger(1944)] Erwin Schrödinger. *What is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, 1944.

[Solomon and Greenberg(2004)] Robert C Solomon and Jeff Greenberg. *Feeling and Thinking: The Role of Affect in Social Cognition*. Cambridge University Press, 2004.

[Sutton and Barto(2018)] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[Tononi(2004)] Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 5(1):42, 2004.

[Tononi(2008)] Giulio Tononi. Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215(3):216–242, 2008.

[Tononi(2016)] Giulio Tononi. *Phi: A Voyage from the Brain to the Soul*. Pantheon Books, 2016.

[Vaswani et al.(2017)] Ashish Vaswani et al. Attention is all you need. *Advances in neural information processing systems*, 2017.

[Wei et al.(2022)Wei, Wang, Schuurmans, Bosma, Chi, Le, and Zhou] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022. URL https://arxiv.org/abs/2201.11903.

[Wikipedia(2021)] Wikipedia. Multi-level feedback queue. https://en.wikipedia.org/wiki/Multi-level_feedback_queue, 2021.

APPENDIX A. OPTOGENETICS — A PROMISING TOOL

Despite decades of research, the mechanisms underlying conscious experience and the transition from unconsciousness to consciousness remain largely unknown. Advances in technology, such as the ability to stimulate and visualize individual neuron cells, may bring us closer to understanding these processes. From a functionalist perspective, it may not be necessary to exactly replicate physical brain operations in order to model consciousness on computers. However, a thorough understanding of these physical mechanisms can provide valuable insights and improve our computer models.

Optogenetics was developed by a team of researchers led by Dr. Karl Deisseroth, a Professor of Bioengineering and of Psychiatry and Behavioral Sciences at Stanford University. The technology was initially described in a series of papers published in the journal Nature in 2005 (Zhang, et al., 2005; Deisseroth, et al., 2005).

Optogenetics involves the use of genetically modified neurons that express light-sensitive proteins called opsins. These opsins can be activated by specific wavelengths of light, allowing researchers to selectively stimulate or inhibit the activity of specific neurons in the brain. Optogenetics has been used in a wide range of studies to investigate the role of specific neurons and neural circuits in various brain functions, including behavior, learning, and memory.

Optogenetics has also been used in a number of clinical studies, including studies of brain disorders such as Parkinson's disease, addiction, and depression. It has the potential to be used as a therapeutic tool for the treatment of these and other brain disorders, although more research is needed to fully understand its potential as a therapeutic intervention.

Optogenetics offers several advantages over traditional electrode-based techniques for studying neural activity. One major advantage is the high spatial and temporal resolution of optogenetic techniques. By using light to stimulate specific neurons, researchers can precisely control the timing and location of neural activity with millisecond precision. This allows researchers to study the function of specific neurons and neural circuits in great detail.

Another advantage of optogenetics is that it allows researchers to study the function of specific neurons in the context of their normal physiological environment. Traditional electrode-based techniques involve physically inserting electrodes into the brain, which can disrupt the normal function of neural circuits. Optogenetics, on the other hand, allows researchers to study neural activity without physically altering the brain tissue.

There are also some limitations to optogenetics. One major limitation is that it can only be used to study neurons that express opsins, which are light-sensitive proteins. This means that optogenetics cannot be used to study the activity of all neurons in the brain, only those that have been genetically modified to express opsins.

Another limitation is that optogenetics requires the use of genetically modified animals, which can be time-consuming and costly to produce. Additionally, optogenetics requires specialized equipment and technical expertise to implement, which can be a barrier to some researchers. Finally, optogenetics is a relatively new technology, and more research is needed to fully understand its potential and limitations.