Modeling Consciousness for System-2 AI: A Review of Theories and Approaches

Edward Y. Chang
Stanford University
echang@cs.stanford.edu

***Abstract*—This paper aims to create a model of consciousness for system-2 AI, which can handle tasks that involve reasoning, planning, and decision making. We examine principles from philosophers and theories from psychiatrists and neuroscientists based on observations from various empirical studies. While we are still uncertain about the true nature of consciousness and how it is situated in the brain, we believe that the implementation of consciousness in machines does not have to strictly follow human anatomy. Our proposed Socrates Computation Model (SCM) 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 specify a case study on a long-term care agent and hope to identify and fix any issues through user studies in future work.**

I. Introduction

System 1 AI, also known as narrow AI, is designed to perform specific, pre-defined tasks. These tasks are typically well-defined and can be executed efficiently by a machine learning algorithm. Examples of system 1 AI include image recognition algorithms and language translation systems. While these types of AI can be very effective at the narrow tasks they are designed for, they are limited in their ability to perform more complex tasks such as reasoning, planning, and decision-making.

To address the limitations of current artificial intelligence systems, 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 the meaning of data and making more informed decisions. The relationship between system 2 and system 1 AI can be compared to the relationship between human consciousness and unconsciousness. While system 1 AI is similar to the unconscious mind, which performs automatic and reflexive actions such as breathing and metabolism, system 2 AI aims to replicate some of the capabilities of human consciousness, which allow us to engage in more complex thought processes.

Studying consciousness is important for many reasons. It is a fundamental aspect of the human experience and plays a central role in how we perceive, think, and interact with the world around us. Additionally, understanding the nature of consciousness and how it arises can help us better understand the complex processes that underlie higher-level thinking and decision-making in the human brain. This knowledge can in turn be used to guide the development of system 2 AI that are capable of similar capabilities.

There are several theories about the nature of consciousness and how it arises, including the Global Workspace Theory, the Integrated Information Theory, and the Dynamic Core Hypothesis. These theories offer different perspectives on the underlying mechanisms of consciousness and the role 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.

The remainder of this article is organized into five chapters towards designing system 2 AI:

- What is consciousness?
- Functionalities of human consciousness,
- Computational model design for consciousness,
- Case study: Design a *knowledgeable, accommodating, loving* long-term care robot, and
- Concluding remarks with open issues and future work.

II. What is Consciousness

Consciousness is the state of being aware of one's own thoughts, feelings, and surroundings II-A. It is a complex and multifaceted concept that has been studied by philosophers, scientists, and theologians for centuries. The precise nature of consciousness and how it arises II-B from the brain and other biological systems is still not fully understood and is a topic of active research. Some theories propose that consciousness is a fundamental aspect of the universe (panpsychism), while others suggest that it emerges from complex computations in the brain (functionalism) II-C.

One reason why we study, and model consciousness is to improve the capabilities of artificial intelligence (AI). Currently, AI known as "system 1" is limited in its flexibility and adaptability. It can only perform narrow tasks quickly and efficiently, but it lacks the ability to adapt to changing environments and unexpected situations (e.g., real-world data distribution drift or out of distribution). This can limit its usefulness in various contexts.

To address these limitations, researchers such as Yoshua Bengio have proposed developing "system 2" AI [Bengio(2020)]. In contrast to system 1 AI, system 2 AI is designed to perform more complex tasks such as reasoning, planning, and

decision-making. It has the ability to understand and interpret the meaning of data, allowing it to make more informed decisions. This type of AI is better suited for tasks that require complex reasoning and decision-making abilities, such as interpretable/explainable natural language processing and autonomous decision-making.

As discussed by Daniel Kahneman in his theory of thinking [Kahneman(2011)], there are two systems of thought: System 1, which is fast and automatic, and System 2, which is slower and more deliberate. While system 1 AI excels at discriminative tasks, system 2 AI excels at generative tasks that require more complex reasoning and decision-making. By studying and modeling consciousness, we can better understand the cognitive processes involved in thinking and decision-making, which can inform the design and development of more advanced AI systems.

In order to model and develop a system that resembles human consciousness, or "system 2," we first need to identify our desired goals and functionalities III. While there have been various theories and hypotheses proposed by researchers in fields such as psychology, philosophy, and theology, these ideas are often abstract and difficult to concretely demonstrate (such as panpsychism vs. functionalism). Instead, we choose to follow Aristotle's first principle of basing our modeling efforts IV on scientific evidence from fields such as physics [Schrödinger(1944)], biology, neuroscience [Deisseroth(2021)], and computer science, rather than relying on more abstract and elusive ideas.

A. Mechanisms of Awareness

Though there has been debates about whether plants and even rocks have consciousness, we focus on modeling consciousness of human capabilities. We consider Michio Kaku's definition on consciousness [Kaku(2014)] to be simple, understandable, and implementable. According to Michio Kaku, the complexity of an organism's consciousness depends on the complexity of its sensing and response system. This means that the more complex an organism's ability to sense and respond to stimuli in its environment, the more complex its consciousness is likely to be.

One way to think about this is to consider the difference between the consciousness of a simple organism like a single-celled amoeba, which has limited sensory capabilities and can only respond to stimuli in very basic ways, and the consciousness of a more complex organism like a human, which has a highly developed nervous system and is able to perceive and respond to a wide range of stimuli in sophisticated ways. The greater complexity of the human brain and nervous system allows us to engage in higher-level thinking and to experience a rich and varied subjective experience.

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

One theory that is similar to Kaku's idea about the relationship between the complexity of an organism's consciousness and its sensory and response system is the Integrated Information Theory (IIT) [Tononi(2004)], which was proposed by Giulio Tononi. This theory argues 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 that it is able to process. The IIT theory is in agreement with Michio Kaku's.

Other theories of consciousness include the Global Workspace Theory, which proposes that consciousness arises from the interaction between different brain areas, and the Dynamic Core Hypothesis, which suggests that consciousness arises from the interaction of different neural networks in the brain. We will discuss these theory when we present *where is consciousness* in Chapter IV-B.

How is consciousness aware of changes in our body and the environment? To explain this, 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 via 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, on the other hand, is an example of top-down processing, which involves conscious planning and execution. In the next section, we will detail the mechanisms behind both top-down and bottom-up awareness.

B. Arise of Consciousness

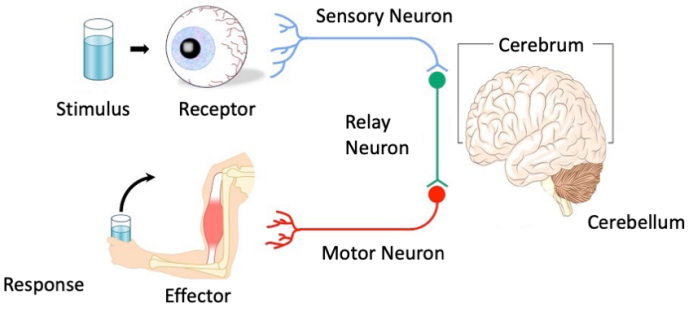

Fig. 1: Bottom-up Attention: Stimulus $\longrightarrow$ Cerebellum $\longrightarrow$ Cerebrum $\longrightarrow$ Response.

In figure 1, the awareness of thirst is informed by unconscious processes and *arises* into consciousness. This process of "arise 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 the stimulus-response diagram, the sensation

of thirst is accumulated unconsciously, and when the signal strength reaches a certain threshold, the thirst experience comes to be experienced or perceived by consciousness. In this case, the unconscious process of accumulating the sensation of thirst is what allows it to arise into consciousness as a conscious experience.

Sigmund Freud was one of the first to propose a model of the mind that included both conscious and unconscious processes [Freud(1900)]. Freud believed that 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. Freud argued that the unconscious mind exerts a powerful influence on our conscious thoughts and behaviors.

Since Freud's time, many other theories have been developed. Carl Jung believed that the collective unconscious is inherited and that it contains the accumulated experiences and knowledge of our ancestors. He argued that the collective unconscious has a powerful influence on our conscious lives and that it can help us better understand the human condition and our place in the world [Jung(1944)]. Unconscious processes play a crucial role in many of the vital functions of the human body, including 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 many other aspects of human behavior and cognition, including memory, perception, and 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 proposes that consciousness is a fundamental aspect of the universe and is present in all matter, including inanimate objects. Some proponents of panpsychism include David Chalmers, Galen Strawson, and Thomas Nagel. This section discusses their philosophical positions and then followed with theories proposed by neuroscientists including Giulio Tononi, Bernard Baars, and Francis Crick, and Christof Koch.

Both Chalmers and Strawson believe that panpsychism may be able to provide a naturalistic explanation for the emergence of consciousness, without invoking supernatural or dualistic explanations. However, their specific views on panpsychism differ in some important ways.

Chalmers is particularly interested in the problem of explaining the subjective nature of consciousness, and he argues that panpsychism may provide a solution to this problem. He also emphasizes the irreducibility of consciousness and the difficulty of explaining it in terms of physical processes.

Strawson, on the other hand, emphasizes the importance of panpsychism as a way of avoiding the *hard problem*[1] of consciousness and as a framework for understanding the nature of the self and its relationship to the physical world. He believes that panpsychism is a more parsimonious and coherent theory of consciousness than other alternatives, such as dualistic or supernatural explanations (e.g., using Hinduism or Buddhism arguments).

Both Chalmers and Strawson agree that panpsychism may be able to provide a framework for understanding the relationship between consciousness and the physical world, and that it may be able to shed light on the nature of consciousness and its place in the universe. Unfortunately, what they provide is a philosophical position and a thinking framework, not a scientific theory and does not have any empirical evidence to support it.

Thomas Nagel is also a philosopher who has made significant contributions to the study of consciousness and its relationship to the physical world. Like Chalmers and Strawson, Nagel is concerned with the problem of explaining the subjective nature of consciousness and the difficulty of reducing it to physical processes. However, his views on this topic differ from those of Chalmers and Strawson in two ways:

- *What it's like*. Nagel argues that subjective experience has an ineffable quality that cannot be fully captured by objective descriptions of the physical processes underlying it.
- *Bat argument*. Nagel has also argued there are limits to our ability to understand the subjective experiences of other beings, even if we have a complete scientific understanding of their physical processes.

Nagel argues that subjective experience has an ineffable quality that cannot be fully captured by objective descriptions of the physical processes underlying it. He believes that the subjective nature of experience is a fundamental feature of the world that cannot be reduced or explained away by any physical theory.

Overall, while both Nagel and Chalmers recognize the difficulty of explaining subjective experience in physical terms, they have different ideas about how this problem might be addressed. Nagel tends to take a more skeptical approach, suggesting that subjective experience may be beyond the reach of scientific explanation. Chalmers, on the other hand, is more optimistic, suggesting that panpsychism may provide a way forward.

[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)].

Panpsychism is a broad and abstract philosophical theory that suggests that consciousness is a fundamental aspect of the universe and is present in all matter, including the brain. This theory has inspired a number of more specific models or theories developed by neuroscientists and psychiatrists in more concrete or detailed ways. Some examples of these derived models or theories include the integrated information theory (IIT) proposed by Giulio Tononi [Tononi(2004)], [Tononi(2008)], [Tononi(2016)], the global workspace theory (GWT) proposed by Bernard Baars [Baars(1988)], the neural correlates of consciousness (NCC) approach proposed by Francis Crick and Christof Koch [Crick and Koch(2003)], [Koch and Tsuchiya(2012)], and the attention schema theory (AST) proposed by Michael Graziano [Graziano(2013a)], [Graziano(2016)]. These models and theories seek to provide more detailed or nuanced explanations for the nature of consciousness, using a range of empirical evidence, including neuroimaging studies and behavioral experiments. They also seek to explore the implications of panpsychism for our understanding of the mind and the physical world. Chapter IV follows up on these models.

$C.2$ 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. 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.

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 computation 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 share the same computation 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.

The functionalism view is a theory of consciousness that proposes that mental states and processes can be fully explained in terms of the functions that they perform. This view is often contrasted with panpsychism and other approaches to the "hard problem" of consciousness [Koch(2004)], 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)]. As we mentioned when discussing panpsychism in Section $C.1$ that both Nagel and Chalmers recognize the difficulty of explaining subjective experience in physical terms. Nagel tends to take a more skeptical approach, suggesting that subjective experience may be beyond the reach of scientific explanation. Chalmers, on the other hand, is more optimistic, suggesting that panpsychism may provide a way forward. We believe that the "hard problem" could be solved by looking into theoretical models that account for these factors and support free will [Dehaene et al.(2003)Dehaene, Sergent, and Changeux].

III. Functionalities of Human Consciousness

There is still ongoing research about the precise nature and functions of human consciousness. This section list a number of key functions, their specifications, together with concerns about programming them into an artificial agent.

Awareness: The ability to be aware of one's surroundings, thoughts, and feelings (Baars, 2002).

In [Baars(1988)], " Bernard Baars presents a theory of consciousness that is based on the idea that consciousness is a global cognitive process that integrates information from various sources and allows an organism to interact with its environment. Baars' theory is centered around the concept of a global workspace, which is a hypothetical system in the brain that allows information from various sources to be integrated and made available to other cognitive processes. According to Baars, consciousness arises when information is broadcast to the global workspace and becomes available to other cognitive processes, allowing the organism to act on it.

Baars' theory also incorporates the idea of attention, which he defines as the process of selecting and focusing on certain pieces of information while ignoring others. According to Baars, attention is a key component of consciousness because it allows the organism to prioritize and act on certain pieces of information while ignoring others.

Researchers have attempted to model the global workspace in several ways, including through computational models, simulations, and brain imaging studies. One example of a computational model of the global workspace is the blackboard architecture, which was developed by Baars and colleagues in the 1980s. The blackboard architecture is a computer simulation that aims 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. Overall, while Baars' theory of consciousness has not been fully validated, it has inspired a significant amount of research and has influenced the development of other theories of consciousness.

Attention: 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.

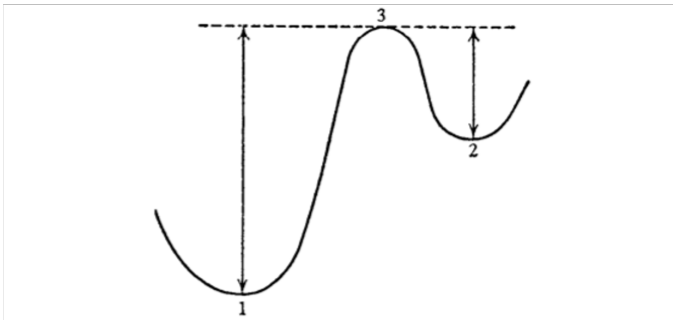

Fig. 12. Energy threshold (3) between the isomeric levels (1) and (2). The arrows indicate the minimum energies required for transition.

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

Erwin Schrödinger's attention model, presented in his lectures collected in the book "What is Life" [Schrödinger(1944)], is consistent with this attention system. Schrödinger describes the bottom-up attention mechanism as working in a manner similar to a "quantum jump" in quantum mechanics. (Figure 2 shows an illustration of a quantum jump in Schrödinger's book.) 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.

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)].

Perception: The process of interpreting sensory information (e.g., via vision) and forming mental representations of the environment [Gregory(1997)].

This area is already adequately supported by system 1 AI, or unconsciousness. The only element that a model ought to consider is how the transitions between unconscious background perception and consciousness awareness is performed. Schrödinger provides the mechanisms in physics to implement the transitions, which is already described under the attention function.

Thinking: 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 the active construction, organization, and evaluation of information, as well as the creation of new ideas and problem-solving.

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

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

There are various theories and models of thinking in psychology that attempt to explain how this process occurs and how it can be influenced by different factors. Some key theories and models include:

- The information processing model, which proposes that thinking involves the manipulation and transformation of information through a series of stages, such as perception, attention, and memory [Miller(1956)].

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

- The cognitive psychology model, which emphasizes the role of mental representations and processes in thinking and problem-solving [Newell and Simon(1972)]. The connectionist model, which proposes that thinking and learning occur through the formation and strengthening of connections between neurons in the brain [Rumelhart and McClelland(1986)].
- The heuristics and biases approach, which suggests that people often use mental shortcuts or rules of thumb to make decisions and solve problems, and that these shortcuts can lead to biases and errors in thinking [Tversky and Kahneman(1974)].
- The dual process model, which proposes that there are two distinct systems of thinking: one that is intuitive and automatic, and another that is more deliberative and controlled [Kahneman(2011)].
- The social cognitive theory, which emphasizes the role of social and cultural factors in shaping thinking and behavior [Bandura(1977)].

These and other theories of thinking in psychology provide important insights into the mental processes involved in generating and manipulating ideas, and can help us understand how thinking can be influenced by different factors and how it can be improved.

There are various approaches to making AI systems that can think and exhibit intelligent behavior. Some common approaches include [Russell and Norvig(2010)], [Laudon and Laudon(2016)]:

- Machine learning: Machine learning involves using algorithms to learn from data and make predictions or decisions. Machine learning algorithms can be trained to perform various tasks, such as image and speech recognition, language translation, and decision-making.
- Cognitive computing: Cognitive computing involves the use of artificial intelligence and machine learning techniques to simulate the way the human brain processes and understands information. This approach aims to replicate human-like thought processes in order to solve complex problems and make decisions.
- Natural language processing: Natural language processing (NLP) is a field of artificial intelligence that focuses on the interaction between computers and humans using natural language. NLP algorithms can understand and generate human-like language, allowing them to communicate with humans in a more natural way.
- Expert systems: Expert systems are AI systems that are designed to replicate the decision-making abilities of a human expert in a particular field. They are built using a combination of rules and machine learning algorithms, and are able to solve complex problems by combining their knowledge and reasoning abilities.

The transformer architecture, introduced in [Vaswani et al.(2017)], allows 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, represented by neighboring entities. This context-based semantic resolution method has significantly improved the performance of natural language processing (NLP) applications. Additionally, 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)].

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

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

Free will and intentionally: The ability to choose goals and to act to achieve those goals [Dennett(1987)].

The debate between free will and determinism is a long-standing and complex one, with strong arguments on both sides. (Figure 3 presents Rembrandt's masterpiece "Adam and Eve" symbolizing this struggle.) While determinism holds

[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.

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 argue that individuals have the ability to make choices and act on them freely (e.g., [Kane(1996)]). 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, such as choosing our words and actions, deciding how to spend our free time, and more seriously, selecting a candidate to vote for.

In the context of AI, the question of whether an AI agent could have free will would depend 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 that an AI system could be designed to make decisions in a more flexible or adaptive way, when the cost of making a bad choice is low. For example, a robot is free to answer a typical question during dating: "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.

However, when the cost is high, such as the debated question of if a self-driving vehicle is about to crash, "should it crash into a barrier to kill passengers or hit a jaywalker?" Since there is no way that a program can in advance be encoded with all possible accident scenarios, an artificial agent will have to make a decision based on some parameters and ethical values (e.g. [Turing(1950)]). If the car has been programmed to prioritize the safety of its passengers above all else, it may choose to crash into the pedestrian. On the other hand, if the car has been programmed to prioritize the safety of pedestrians above all else, it may choose to crash into the barrier. In either case, the decision made by the car would still be based on the specific criteria and rules that it has been programmed with, rather than on any sense of free will. Note that a random decision is not considered a decision made of free will.

So, what are the specific criteria and rules that an AI agent has to be programmed with? It depends on the specific task or problem that the AI is designed to solve. In general, when the cost of a bad choice is not negligible, considerations of laws and ethics during decision making are essential. The higher the stake of a decision, the lower the risk an agent can take. Furthermore, since laws and ethics are not universal in all regions of the world, an agent must be able to adapt to the "environment" as it moves. In general, we can say that there are external rewards and penalties that an agent must consider when making a decision. In addition, human beings have internal awards that must be considered. For example, some may not tolerate a long-term care robot to tell a joke using bad words. Therefore, during the process of making a choice, we are influenced by our intentionality and so should an AI agent.

Intentionality is the capacity of an individual or system to have mental states, such as beliefs, desires, and intentions, that are directed towards objects or states of affairs in the external world. It is a framework for understanding and predicting the behavior of complex systems, including other people and the environment. D. C. Dennett has written extensively about intentionality and its role in shaping human behavior [Dennett(1987)].

Free will and intentionality are closely related concepts, but they are not the same thing. Free will is about the ability to make choices and act freely, while intentionality is about the way that the mind is directed towards or about something. To illustrate the differences between free will and intentionality, consider the following example: Sarah is trying to decide what to have for breakfast. She has the choice between having eggs or cereal. She weighs the pros and cons of each option and ultimately decides to have eggs. In this case, Sarah's decision to have eggs is an example of free will, as she had the ability to choose between two options and made a decision based on her own preferences and desires. The intentionality in this situation is the mental state of wanting to have a satisfying breakfast, which directed Sarah's decision-making process towards considering different options and ultimately choosing eggs. Intentionality can be coded as intrinsic rewards that we have mentioned.

Individuals may be motivated by a desire to maximize their total intrinsic reward or utility from their choices. This reward function may include a variety of factors, such as happiness, pleasure, satisfaction, and other goals and values. The specific factors that are included in an individual's reward function are likely to be influenced by a variety of factors, such as their personal experiences, upbringing, education, cultural background, and other factors.

To optimize total rewards, both internal and external, individuals may consider using a variety of quantifiable measures to evaluate the potential outcomes of their choices. For example, they may use cost-benefit analysis to compare the costs and benefits of different options, or they may use decision-making tools such as decision trees or utility functions to help them weigh the trade-offs between different considerations. Other techniques that may be helpful in optimizing total rewards include goal-setting and prioritization, risk assessment, and contingency planning.

It is worth noting that the pursuit of total reward or utility is not always straightforward, and individuals may need to make trade-offs between different factors in their reward function. For example, they may need to balance their desire for immediate pleasure or enjoyment with their long-term goals or responsibilities. Additionally, the relative importance of different factors in an individual's reward function may change over time or in different circumstances, and individuals may need to adjust their approach to optimizing total rewards accordingly.

Regarding implementation, free will can be formulated by the values of choices and the entropy among choices [Rehn(2022)].

This means that an individual's free will is represented by the various options they have to choose from and the inherent uncertainty or randomness in their decision-making process.

Emotion: The experience of feelings and emotional states, and the ability to express and respond to emotions.

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 Model Design

We have this far presented the theories of consciousness, panpsychism and functionalism, in Section II-C, and key functionalities of human consciousness that are useful to support by machines that implements system 2 AI in Section III. This section grounds these theories and functionalities into architecture and design specifications. We first present theoretical models and then a proposal of a computational model.

A. Review of Theoretical Models

Let us first discuss representative theoretical models developed by psychiatrists and neuroscientists. Their designs are founded on brain study and behavior study, rather than just some philosophical thoughts. These models attempt to explain how consciousness arises from the activity of the brain, and how it might be simulated in artificial systems such as computers.

Global Workspace Theory (GWT): GWT was proposed by Bernard Baars in 1988 [Baars(1988)]. GWT is also known as the *global neuronal workspace* (GNW). Baars' contributions to the development of the GNW theory include the following:

1) He developed the original conceptual framework for the GNW theory, which proposes that consciousness arises from the integration of information from multiple brain systems through interactions in the prefrontal cortex.
2) He proposed the idea that the prefrontal cortex acts as a central hub that integrates information from various brain systems and broadcasts this information to other brain regions, a key aspect of the GNW theory.
3) He developed a computational model of the GNW theory, which specifies the algorithms and processes that underlie the integration of information in the prefrontal cortex.
4) He provided empirical evidence for the GNW theory through a series of experimental studies, which have been influential in shaping our understanding of the neural basis of consciousness.

Integrated Information Theory (IIT): According to IIT, proposed by Giulio Tononi [Tononi(2004)], consciousness arises from the integration of information across multiple sources and is characterized by the ability to distinguish between different states of the system. IIT proposes that the

amount of consciousness in a system can be quantified by a measure called ϕ, which reflects the degree of integration of information in the system. For example, consider a simple system with two parts, A and B. If the state of A does not affect the state of B, and the state of B does not affect the state of A, then there is no integration of information between the two parts, and the system has zero ϕ. On the other hand, if the state of A does affect the state of B, and the state of B does affect the state of A, then there is integration of information between the two parts, and the system has non-zero ϕ value.

Dynamic Core Hypothesis (DCH): DCH was developed by neuroscientist Bernard Edelman [Edelman and Tononi(2000)]. DCH proposes that consciousness arises from the activity of a specialized region of the brain called the thalamocortical system, which is made up of the thalamus and the cortex. According to the theory, the thalamocortical system is a dynamic network of neurons that constantly adjusts its activity in response to sensory input, motor output, and other internal and external influences. This dynamic activity is thought to give rise to the subjective experience of consciousness.

One key aspect of the Dynamic Core Hypothesis is the idea that consciousness is not a unitary, all-or-none phenomenon, but rather a continuous, graded property of the brain. According to the theory, different brain states can give rise to different levels of consciousness, depending on the degree of integration and differentiation of activity in the thalamocortical system. For example, a person who is deeply asleep or under general anesthesia would have a low level of consciousness, while a person who is fully awake and alert would have a high level of consciousness.

The Dynamic Core Hypothesis also proposes that consciousness is closely tied to the brain's ability to generate predictions about the environment and to update those predictions based on new information. This predictive coding process is thought to be mediated by the thalamocortical system, and it is thought to be essential for the experience of consciousness.

Attention Schema Theory (AST): AST is a theory of consciousness developed by neuroscientist Michael Graziano [Graziano(2013b)], [Graziano(2016)]. AST proposes that consciousness arises from the brain's representation of attention as a computational process. According to the theory, the brain constructs a model or schema of attention that it uses to represent the attention states of both itself and others. The attention schema for itself creates a sense of self and a subjective experience of attention. (This attention schema is thought to be a way of representing the concept of attention in the brain, rather than being a direct representation of attention itself.)

According to AST, the attention schema is constructed by the brain in order to facilitate communication and social interaction. It allows the brain to predict and interpret the attention states of others, which is important for understanding their intentions and behaviors. The attention schema is also thought to play a role in the brain's ability to allocate its own attention resources, allowing it to focus on relevant stimuli and ignore distractions.

Remarks:

These four models are mutually consistent. The key takeaways are consciousness deals with data collection, information exchange, and integration. It pays attention to what is considered to be more important, and the level of attention can be measured and quantified. Together with long-term memory, consciousness conducts *subjective* understanding, prediction, and decision making. The factors/parameters that these theoretical models do not consider include rewards, ethics (internal rewards), and the interactions between consciousness and subconsciousness. From a computer scientist's perspective, these theoretical models do not differentiate foreground vs. background processing, not do they consider motivating/prioritizing conscious activities by external and intrinsic rewards.

B. 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.

C. Consciousness and Unconsciousness Transitions

There are several theories about the mechanisms underlying the transition from unconsciousness to consciousness. We have mentioned that GWT considers that the prefrontal cortex acts as a central hub for integrating and processing information from various sensory and cognitive systems. When an event or stimulus is perceived as important or novel, it is thought to be broadcast to the prefrontal cortex and other brain regions for further processing and action. This process is thought to give rise from unconscious to conscious experience with respect to the event. Similarly, The IIT theory suggests that consciousness arises from the integration of information across multiple brain systems. According to this theory, the transition from unconsciousness to consciousness involves the integration of information in a way that is both rich and differentiated.

Erwin Schrödinger, a famous physicist, put the theories and empirical evidences into a quantifiable framework that computers can execute. Schrödinger discussed the concept of the transition from unconsciousness to consciousness in his book "What is Life?". Schrödinger proposed that the second law of thermodynamics, which states that entropy (a measure of disorder or randomness) in a closed system will always increase over time, could be used to understand the transition from consciousness to unconsciousness. He suggested that this transition occurs when there is an increase in the randomness or disorder of matter, such as when a task becomes habitual and the brain enters an inert state.

In addition to the second law of thermodynamics, Schrödinger also proposed that quantum mechanical processes, such as quantum jumps or superpositions, may play a role in the transitions between unconsciousness and consciousness. He suggested that these processes may be triggered when the accumulated energy in a neural region reaches a threshold. When this happens, the region of the brain responsible for managing consciousness is alerted to pay attention to the emergent event and to plan for reactions.

Overall, while the exact mechanisms underlying the transition from unconsciousness to consciousness are not fully understood, the global workspace theory and the integrated information theory provide important insights into this process. Schrödinger's ideas also suggest that the second law of thermodynamics and quantum mechanical processes may be involved in these transitions.

D. 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.

E. Computational Model

We propose the Socrates Computation Model (SCM), which consists of four logical modules: the receptor, unconsciousness, consciousness, and effector modules.

- 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 act as a receptor and sends 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.

Scheduler: In a multi-level feedback queue (MFQ) scheduler [Wikipedia(2021)], processes/tasks are organized into a hierarchy of queues, with each queue having a different priority level. The scheduler selects the process at the front of the highest priority queue that has processes in it. If a process uses up its time slice without finishing, it is moved to the back of the queue at its current priority level. If a process finishes, it is removed from the queue.

A MFQ scheduler can be configured with different time slices (also known as quantum sizes) for each queue, with shorter time slices for higher priority queues and longer time slices for lower priority queues. This allows processes in higher priority queues to be serviced more frequently, while giving processes in lower priority queues the opportunity to run if the higher priority queues are empty.

MFQ scheduling can be used to ensure that important processes receive a higher priority and are serviced more quickly, while also allowing lower priority processes to make progress if the system is not overloaded. This can improve the overall responsiveness and fairness of the operating system.

Priories are decided by a task's value and importance, which can be quantified by external and internal rewards, which discuss next. For instance, an event that can relieve pain may be considered to be high in priority to be attended to by consciousness.

External reward system for adaptivity:

Using rewards to train an AI agent to behave in an optimal way to 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.

There are also other approaches that can be used to make an AI agent more adaptive to its users and environment. One such approach is learning from demonstrations, where the agent is trained to imitate the actions of a human expert or teacher. This can be an effective way to transfer knowledge and skills from humans to the AI agent, and can be especially useful when it is difficult to explicitly specify the desired behavior of the agent in terms of a reward function. With recent advances in large pre-trained language models (LLMs), one can provide demonstrations via a list of examples through prompts. Prompts can be regarded as a template, which starts with a description of the goal, and then some specific and focused instructions with a list of examples, the more the better. In summer 2022 at Stanford, we initiated the *Noora* chatbot project, which aims to help autism patients to learn how to speak empathetically. 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."

Example #1:"I was laid off by my company today!"
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."

One way to improve Noora's adaptability is to provide feedback on her responses. For example, if the user is dissatisfied with a non-empathetic response, they can give feedback indicating that an empathetic response would acknowledge their feelings and offer support and understanding, rather than providing abstract or cold advice. 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.

Intrinsic reward system for ethics:

Intrinsic rewards can be used in reinforcement learning to shape the behavior of an agent in an ethical manner. These rewards can be provided as positive reinforcement for actions that align with ethical values and principles. By acting ethically, the agent derives satisfaction and fulfillment from making its users satisfied and fulfilled.

Similarly to how we model external rewards, the intrinsic reward system can use reinforcement learning and prompting templates to teach desired behaviors and ethics to an AI agent through demonstrations. For example, the template for empathy can be used to model other positive behaviors such as being attentive, caring, and humorous (see some positive emotions listed in Section III). It is important to note that a machine may already possess some positive characteristics, but we still need to explicitly model good and bad behaviors for the agent to understand and work effectively 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 made adaptive to the individual and their local culture and laws.

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

As engineers and researchers, we recognize that there are several key challenges facing senior care and long-term care in the United States and beyond, including issues of affordability, 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 are considering the development of an artificial agent that could assist with various day-to-day tasks. In this case study, the tasks are to:

1) Identify key requirements for a healthcare artificial agent.

2) Conduct a feasibility assessment on the top five requirements on the list.
3) Identify three requirements that could potentially be addressed by existing system-1 AI capabilities, and discuss how these could be implemented.
4) Choose one "hard" requirement that we plan to tackle through the development of system-2 AI capabilities.
5) Sketch out our design and technical tasks, and work towards implementation.
6) Make the design adaptive to personal preferences.

At the top, abstract level, we would like an agent to be able to make a care-receiver a better version of herself day after day. How "better" is quantified is individual-dependent, and can be understood by the robot in time. Examples of feeling better can be healthier, happier, and "beautifuler". Suppose that we desire a robot to be attentive, loving, and humours. How should these characters or attributes be quantified? All methods for implementing a working prototype have been presented in the previous sections.

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 Socrates Computation Model (SCM), 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.

In addition to these challenges, we must consider the issue of free will and how it can be accurately modeled within an AI system. AI agent to not only understand and predict its own states, but also the states of its users and the surrounding environment in order to effectively interact.

One challenge that has been identified is the need for AI agents to not only understand and predict their own states, but also the states of their users and the surrounding environment in order to effectively interact. Another challenge is determining whether free will should be allowed and how freedom can be accurately modeled. To demonstrate the potential of SCM, a case study has been provided on the development of a long-term care agent with a list of requirements. It is believed that SCM will enable the creation of a prototype, and through user studies, any shortcomings and open issues will be identified and addressed in future work.

Acknowledgment

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 not only keep track of its self-state but also the users' state.
- Providing the bibtex format of the references.

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(2022)] Edward Y. Chang. Towards artificial general intelligence via consciousness modeling (invited talk). In *IEEE Infrastructure Conference*, September 2022. 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.

[Edelman and Tononi(2000)] Gerald M Edelman and Giulio Tononi. A dynamic core model of conscious and unconscious brain functions. *Proceedings of the National Academy of Sciences*, 97(2):1944–1953, 2000.

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

[Graziano(2013b)] Michael S.A. Graziano. Attention schema theory: A novel theory of consciousness. *Frontiers in psychology*, 4:1–11, 2013b.

[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.

[Turing(1950)] Alan Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[Tversky and Kahneman(1974)] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[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.