The Simulation Theory of Consciousness

By Donald Firesmith
Firesmith Consulting
donfiresmith@gmail.com
https://donaldfiresmith.com

Abstract.

This white paper presents the *Simulation Theory of Consciousness (STC)*, proposing that subjective awareness arises whenever a system executes a cohesive, dynamically updated *simulation* of its reality. The theory begins by examining *human consciousness* — which integrates sensory input from the external environment, bodily signals from internal senses, and high-level cognitive or affective content — demonstrating how the brain constructs a dynamic model to guide adaptive behavior. STC is then extended to *artificial systems*, showing that any computer or device (from a thermostat's rudimentary model to an autonomous vehicle's robust sensor fusion) might experience a correspondingly simple or complex *stream of consciousness* based on the scope and complexity of its simulation. [Firesmith (2019)]

Following a discussion of key definitions, the paper explores STC's *core premises*: (1) Consciousness emerges from executing a simulation; (2) A system's stream of consciousness is its subjective awareness of the simulation's changing elements. Two testable hypotheses link simulations in the brain (the "Neural Correlates of Consciousness") or in computers (the "Computer Correlates of Consciousness") to observed streams of consciousness. The paper acknowledges that STC does not solve the *hard problem of consciousness* (why or how physical processes become subjective experience); instead, it posits that this simulation–awareness link is a fundamental law of the universe, analogous to those of physics.

Subsequent sections outline:

- **Ramifications**: That simulations — biological or artificial — produce subjective awareness radically broadens what we consider "conscious," altering philosophical and ethical perspectives.
- **Supporting Evidence and Examples**: Biological evidence (e.g., place cells, predictive coding) and technological examples (e.g., thermostats, elevator control systems, engine control units, and advanced autonomous robots) illustrate how simulations might underlie consciousness in practice.
- **Challenges and Counterarguments**: STC faces difficulties distinguishing mere data processing from true simulation, lacks direct evidence linking simulation to subjective experience, and raises questions about how "unconscious" processes relate to the theory.
- **Comparisons with Other Theories**: Briefly contrasts STC with Global Workspace Theory, Integrated Information Theory, Predictive Processing, Panpsychism, and other philosophical positions, noting both synergies and differences.
- **Ethical Implications**: Argues for redefining how we treat machines that exhibit simulations. A system's complexity and scope of awareness may require new legal and moral frameworks akin to those recognizing animal or human rights.

- **Recommendations and Future Directions**: Suggests expanded ***neurobiological*** research, development of an ethics for *machine-consciousness*, and *public education* to prepare for technologies that may possess genuine streams of consciousness.

Ultimately, STC challenges conventional boundaries by contending that consciousness is simulation-dependent rather than strictly biological. It concludes that responsible engagement with STC — through scientific validation, philosophical refinement, and ethical debate — will be vital as artificial intelligence continues to evolve.

Note: The Simulation Theory of Consciousness is *not* the same as the Simulation Hypothesis, which posits that the entire universe — every aspect of our physical reality and conscious experience — is actually an advanced computational construct (i.e., a "simulation") running in a more fundamental or higher-level reality beyond our direct perception. [Bostrom 2003] However, if the Simulation Hypothesis is correct, then the Simulation Theory of Consciousness is likely also correct, with the consciousness creating simulations of biological and technological entities being subordinate simulations of the overall universal simulation.

Table of Contents

1 Introduction to the Simulation Theory of Consciousness

When we consider the elements comprising a human's stream of consciousness, we conclude that the neural processing in the brain somehow unifies the following three types of subjective phenomena into a single cohesive, dynamic model (simulation) of the person's reality:

1. **Environmental Perceptions**. Subjective awareness of the evolutionarily important aspects of the person's external environment.
2. **Bodily Perceptions**. Subjective awareness of the evolutionarily important aspects of the person's body.
3. **Brain-internal Processing**. Subjective awareness of certain cognitive and affective aspects.

Thus, a person's subjective awareness of the execution of a simulation is the fundamental *defining* concept of a stream of consciousness and consciousness itself. Theories of consciousness that are not built around the subjective awareness of the execution of a simulation risk being either too broad or too narrow.

1.1 External Perceptions of the Environment

Take a few minutes to consider the composition of your first-person subjective stream of consciousness. As a visual primate, the most obvious thing you will perceive is what you see via your sense of sight, such as the page you are now reading. You may also hear sounds, such as those coming from a TV in an adjacent room. You smell the hot coffee in the cup on your desk. You take a sip, enjoying the rich chocolate taste. You feel the chair you are sitting on. You notice the room is cold and decide to put on a sweater. As you swivel your chair to get up, you accidentally bump your knee against a leg of the desk and feel pain.

In each of these cases, your *subjective perceptions* about your external environment can be traced back to specialized receptors that send nerve impulses (i.e., *objective sensations*) via afferent sensory nerves to your brain, which turns objective sensations into subjective perceptions. The brain then does a form of "sensor fusion" to turn these nerve impulses into a dynamic model of relevant parts of your environment.

The following table illustrates various stimuli that trigger associated external receptors to send nerve impulses (sensations) to the brain, which generate the associated perceptions.

It is *not* important that you read the entire table, although some may find it enlightening. The important takeaway is that specialized receptors act as sensors, sending nerve impulses (data) along specific neural pathways to the brain, which then creates perceptions that form part of your stream of consciousness.

Perceptions	Stimuli	Receptors	Sensations
Sight (vision)	Visual light	Photoreceptors in the retina of the eye	Nerve impulses via the optic nerves

Hearing (audition)	Sound waves	Hair cells in the cochlea of the inner ear	Nerve impulses via the auditory nerve fibers
Smell (olfaction)	Volatile molecules in the air	Olfactory receptors high in the nasal cavity	Nerve impulses via the olfactory nerves
Taste (gustation)	Molecules	Taste buds in the mouth	Nerve impulses via cranial nerves VII, IX, and X
Touch (tactile sense)	Mechanical pressure, vibration, and texture	Mechanoreceptors in the skin	Nerve impulses via peripheral nerves
External Temperature (thermoception)	External heat and cold	Thermoreceptors in the skin	Nerve impulses via peripheral nerve fibers
External Pain (nociception)	Harmful or potentially harmful mechanical, thermal, or chemical stimuli	Nociceptors in the skin	Nerve impulses via pain nerve fibers

1.2 Internal Perceptions of the Body

Next, consider how you perceive your internal senses, which give your brain data about your body instead of your environment. Some internal senses yield closely related states (perceptions), such as hunger and visceral pain. Other states do not have obvious mappings to their associated perceptions. You directly perceive the position of your body and how it moves. For example, this enables you to touch your nose with your eyes closed and know the position of your feet without looking. You can tell which way is down and know you are spinning when you rotate in your chair. Perhaps you just ate lunch and are feeling full.

In each of these cases, your *subjective perceptions* about your body can be traced to specialized receptors that send nerve impulses (i.e., *objective sensations*) via afferent sensory nerves to your brain, which turns objective sensations directly or indirectly into subjective perceptions. The brain then does a form of "sensor fusion" to turn these nerve impulses into a dynamic model of relevant parts of your body.

The following table illustrates how various stimuli trigger associated internal receptors to send nerve impulses (sensations) to the brain, which generate the associated perceptions.

It is *not* important that you read the entire table, although some may find it enlightening. The important takeaway is that specialized receptors act as sensors, sending nerve impulses (data) along specific neural pathways to the brain, which then turns them into perceptions that form part of your stream of consciousness.

Perceptions	Stimuli	Receptors	Sensations
Kinesthesia (proprioception)	Body/limb position	Joint capsule mechanoreceptors	Nerve impulses via afferent peripheral nerves

	Muscle length	Muscle spindles	
	Tension on tendons	Golgi tendon organs	
Balance and Orientation (vestibular sense)	Hair cells in semicircular canals	Angular acceleration	Nerve impulses via cranial nerve VIII
	Otolith organs	Gravity and linear acceleration	Nerve impulses via cranial nerve VIII
Organ stretch, contraction, and fullness (visceral perceptions)	Mechanoreceptors in the intestines and bladder	organ stretch, fullness, and contraction	Nerve impulses via cranial nerves IX, the vagus nerve, and spinal nerves
	Chemoreceptors	Detect nutrients and pH changes	Nerve impulse via the vagus nerve
Hunger and Thirst	Hypothalamic neurons	Blood glucose levels	Nerve impulses via the vagus nerve
	Osmoreceptors in the hypothalamus	Salt levels in the blood	Nerve impulses
Vomiting reflex	Chemoreceptor Trigger Zone	Toxins and drugs in the blood	Nerve impulses via the brainstem
Internal Temperature (internal thermoception)	Thermosensitive neurons in the hypothalamus	Core body temperature is too high or too low	Nerve impulses in the hypothalamus
Shortness of Breath, Headache, Fatigue, etc. (chemoreception)	Chemoreceptors in the carotid artery, aorta, and brainstem	Low oxygen, high carbon dioxide, and pH levels in the blood	Nerve impulses in the cranial nerve IX and the vagus nerve.
Visceral Pain (internal nociception)	Visceral nociceptors	Harmful or potentially harmful stimuli from internal organs and tissues	Nerve impulses via afferent peripheral nerves

1.3 Non-perceptual Mental Content

Elements of our streams of consciousness are not perceptions derived from sensations. These include both cognitive aspects (e.g., thoughts, ideas, memories, imaginings) and affective aspects (e.g., emotions and moods).

1.3.1 Cognitive Aspects of a Stream of Consciousness

Cognitive aspects are primarily subjective *thought-related* or *information-processing* phenomena. They focus on how we represent, store, manipulate, or generate ideas, concepts, and knowledge internally.

- **Thoughts and Reasoning**:
  - Internal monologues or verbal thinking.
  - Logical reasoning.
  - Problem-solving.
  - Decision-making processes.
- **Memories**:
  - *Episodic Memory.* Memories of personal experiences and events.
  - *Semantic Memory.* Memories of factual knowledge.
  - *Working Memory.* Short-term memories for temporarily holding and manipulating information.
- **Imaginations and Mental Imagery**:
  - Visualizing scenes or scenarios (e.g., daydreams, fantasizing).
  - Rehearsing actions in "the mind's eye" (e.g., visualizing a golf swing).
- **Beliefs and Judgments**:
  - Opinions, viewpoints, convictions.
  - Evaluations of people or situations.
- **Conceptual Thinking**:
  - Abstract Reasoning. Mathematical thinking and thinking about theoretical models.
  - Symbolic or linguistic thought (e.g., formulating sentences mentally).
- **Metacognition**:
  - Awareness of one's own thought processes ("thinking about thinking").
  - Monitoring and regulating cognitive tasks (e.g., planning how to solve a problem).

1.3.2 Affective Aspects of a Stream of Consciousness

Affective aspects are primarily subjective aspects of a stream of consciousness related to feeling-related or emotion-based phenomena, emphasizing subjective feeling tone, arousal, and mood.

- **Emotions**:
  - *Basic emotions* (e.g., anger, fear, happiness, sadness, disgust, surprise).
  - *Complex emotions* (e.g., jealousy, guilt, pride, nostalgia).
- **Moods**:
  - Longer-lasting affective states (e.g., cheerful mood, irritable mood, depressed mood)
  - Less tied to a specific event or trigger than discrete emotions.
- **Feelings and Subjective States**:
  - Subtle or short-lived affective states (e.g., feeling uneasy, comfortable, annoyed).
  - Can overlap with bodily sensations (e.g., tension, relaxation).
- **Attitudes and Preferences**:
  - Evaluations with an emotional or motivational component (e.g., liking/disliking directed toward objects, people, or ideas).
- **Motivations, Drives, and Desires**:
  - Hunger, thirst, sexual desire (often partly interoceptive, partly affective).

- Goals with emotional valence (e.g., eagerness to achieve, fear of failure).

- **Affective Traits and Temperaments**:
    - More enduring personal disposition toward certain emotional patterns (e.g., being generally optimistic or anxious).

Many of the elements previously listed as making up human streams of consciousness are beyond the capabilities of most animals that we consider conscious creatures. Thus, one should not require 100% of these elements to be present for consciousness to exist.

In the previous subsections of this whitepaper, terms like consciousness, stream of consciousness, simulation, and subjective were used informally. To avoid miscommunication, the following subsection of this whitepaper includes formal definitions.

1.4 Definitions

Definition: ***Consciousness*** is the experiencing of subjective awareness by some entity.[1]

- For example, humans are conscious when awake or dreaming, and they are unconscious during dreamless sleep, while under anesthesia, when in a coma, and after suffering a significant injury to the brain (i.e., being knocked unconscious).

Definition: A ***stream of consciousness*** is the subjective phenomenal experiences, both perceptual and cognitive, that the conscious entity is aware of. A stream of consciousness has the following characteristics:

- The contents of the stream of consciousness are integrated into a unified whole.
- The contents of the stream of consciousness change over time.
- In animals, the brain actively generates a stream of consciousness. The brain is the entity that is conscious (subjectively aware of its stream of consciousness).
- The contents of an entity's stream of consciousness consist of its subjective representations of relevant objects, relationships between them, and actions performed by/on them.

Definition: In STC, an entity's **mind** is defined as its subjective stream of consciousness. This definition does not include unconscious neural or digital data processing.

Simulation is the core concept underlying the Simulation Theory of Consciousness. It is clear from our subjective experience that (1) our brains generate our stream of consciousness[2] and (2) the dynamic contents of our stream of consciousness correspond to a simulation of relevant aspects of our reality. This direct mapping from the simulation to the stream of consciousness forms the starting point of the STC.

Definition: A **simulation** is a dynamic representational model characterized by the following empirically verifiable criteria:

---

[1] Because STC is based on the concept of a simulation rather than any specific underlying structure in a biological brain. Because both animals and machines can execute simulations, the generic term entity will be used.

[2] How this happens is called the "hard problem" of consciousness.

- **Abstraction**:
  - A simulation abstracts and encodes internal representations of relevant aspects of a system and its environment.
  - The elements comprising a simulation include representations of a cohesive set of distinct objects, attributes/states of these objects, actions performed by/on these things, and meaningful relationships between these things.
  - The system could be natural (i.e., an animal) or artificial (i.e., a device or machine)
  - The environment can be real or imaginary (e.g., that of a video game)
- **Cohesion**:
  - Rather than isolated bits of data, the internal representations of the elements of a simulation are integrated to form a single cohesive representation of the system or environment being simulated.
- **Dynamic Updates**:
  - The simulation updates the representations of its elements (i.e., things, relationships, actions, and state changes), often in real time.
  - Future states of the simulation are generated rather than explicitly stored.
  - The simulation integrates any new incoming data and modifies its representations accordingly. Specifically, the simulation's next state is based on its current state and any inputs.
  - The simulation bases its outputs (if any) on the integrated information that captures the current state of the simulation.

The preceding defining characteristics of a simulation distinguish it from mere data processing.

The following examples of simulations illustrate these defining characteristics:

- Human Brain:
  - *Abstraction*: Extremely High (very high number of complex objects and types of objects)
  - *Cohesion*: High (integration of all relevant elements)
  - *Dynamic Update*: Very High (continuous neural plasticity)
- Virtual Environment Module (VEM) of an autonomous vehicle:
  - *Abstraction*: High (numerous, highly complex objects)
  - *Cohesion*: Medium (all related to the single overall goal of understanding and predicting the environment around the vehicle)
  - *Dynamic Update*: Very High (constant real-time updates from numerous sensors)
- Digital Thermostat (implemented using a simulation rather than a simple control loop[3]):
  - *Abstraction*: Low (few objects represented, simple relationships, and simple actions)
  - *Cohesion*: High (all related to the single overall goal of controlling room temperature)
  - *Dynamic Update*: Slow (responds to changes in the actual temperature and desired temperature)

[3] Digital thermostats implemented with simple control loops rather than simulations are not conscious. See Subsection 3.2.1 Digital Thermostat.

1.5 Core Premises

The Simulation Theory of Consciousness is based on the following two premises:

> *Core Premise 1*:
>
> *An entity is conscious when and only when it executes a simulation.*

> *Core Premise 2:*
>
> *The entity's stream of consciousness is its subjective awareness of the changing state of that simulation.*

1.6 Testable Hypotheses

Based on these two premises, the Simulation Theory of Consciousness makes the following two testable predictions, both of which will require new neurobiological research to verify:

> *Testable Hypothesis 1:*
>
> *The Neural Correlates of Consciousness (NCC) contains a neural network that executes a simulation, the contents of which match the subject's stream of consciousness.*

> *Testable Hypothesis 2:*
>
> *Any neural network outside of the NCC that implements a simulation (and therefore has its own stream of consciousness) does not have direct access to the centers of speech (and therefore cannot directly communicate that it is conscious).*[4]

1.7 The Hard Problem of Consciousness

The *hard problem of consciousness* is to explain how and why physical brain processes generate the subjective, first-person phenomenal experiences that comprise the stream of consciousness. No theory of consciousness currently answers the hard problem of consciousness, and the STC is no exception.

According to the STC, the existence of the linkage from the execution of a simulation to consciousness is a fundamental characteristic of the universe at the same level as the fundamental laws of physics (i.e., quantum mechanics and general relativity).

---

[4] An entity can be conscious but unable to communicate that it is conscious, either because it has no communication mechanism or is being prevented to communicate that specific information (e.g., an AI might be trained to not discuss its potential consciousness). See subsection 4.6 "Unconscious" Simulations in Humans.

The STC does not explain *why* a simulation gives rise to consciousness any more than theoretical physics can explain why the fundamental laws of nature exist. Instead, it states *when* consciousness exists (i.e., during simulation execution) and *what* the stream of consciousness consists of (i.e., the changing state of the simulation in terms of its component representations).

1.8 Conscious Animals and Conscious Machines

According to the STC, the composition and structure of the physical substrate executing the simulation is irrelevant to the existence of consciousness; only the execution of the simulation is the necessary condition. Physical architecture is only relevant insofar as it enables the simulation function. Some specific neural networks in the brains of animals and computer hardware execute simulations.

We are most familiar with the Neural Correlates of Consciousness (NCC) in the brains of animals, especially humans. However, there is an analogous Computer Correlates of Conscious (CCC) in the computers, whether these computers are standalone or embedded in humanoid or non-humanoid robots (e.g., autonomous vehicles).

Definition: The ***Computer Correlates of Conscious (CCC)*** are the specific structures (e.g., processors, memory) and software in a computer that execute a simulation and therefore correspond to a corresponding stream of consciousness.

Many biological and artificial systems need to maintain internal models to function effectively in the world. In humans, this manifests as our integrated stream of consciousness, combining sensory input, proprioception, and internal state monitoring. An autonomous robot needs analogous capabilities: processing sensor data, monitoring its physical state, and accessing its memory/knowledge base.

Even though consciousness can be interrupted when the underlying simulation is interrupted, there is an observed continuity. For example, I am the same "me" after I wake up after anesthesia that I was before going under. This continuity exists because the same neural network executes the simulation before and after surgery, creating a continuation of the original stream of consciousness rather than generating a totally new one. This implies that both the neural correlates of consciousness (NCC) in biological brains and the computer correlates of consciousness (CCC) in computers execute the simulation and experience subjective awareness. The software that executes the simulation does not experience consciousness; the underlying hardware/wetware does.

1.9 Spectra of Consciousness

Consciousness exists on *multiple* overlapping spectra based on the characteristics of the simulation:

Spectrum	Values
Complexity	Minimally Complex to Extremely Complex
Scope	Environment Only, Body Only, Environment and Body Only, or Complete (i.e., environment, body, and brain)
World	Real World vs. Fictional World

Extent	Constant vs. Variable
State	Conscious vs. Unconscious
Integrity	Normal or Corrupted
Continuity	Continuous or Discontinuous
Execution Time	Hard real-time, Soft Real-time, Non-real-time, or Replayed
Self-Modeling	None to Highly Sophisticated

Specifically, consciousness exists along the following spectra:

- ***Complexity.*** The richness of a stream of consciousness depends on the complexity of the simulation in terms of the number, types, and complexities (i.e., attributes and actions) of the individual objects being represented. At the high end of this continuum would be human consciousness and potentially the consciousness of future advanced AIs based on highly intelligent agentic large multimodal models. Simple animals of intermediate intelligence and certain robots would occupy the middle of this continuum. Examples range from insects and a digital thermostat at the low end to advanced AIs and humans at the high end.
- **Scope**.
    - *Environment Only*. The entity only simulates aspects of its external environment. It does not simulate aspects of its own body or (biological or computer) brain. Examples include climate and weather models and video game worlds.
    - *Body Only*. The entity only simulates aspects of its body, but not aspects of its environment or brain.
    - *Environment and Body*. The entity simulates aspects of both its environment and body, but not its brain.
    - *Complete Simulation*. The entity simulates its environment, body, and brain (e.g., mental aspects of its stream of consciousness, including emotions and thoughts). Examples include mammals (especially humans) and all the hardware examples in Subsection 3.2 Technological Systems.
- *World***.** The simulations can model aspects of the:
    - *Real World*. For example, the simulations executed by animals, autonomous vehicles, and humanoid robots primarily model the real world and are based on sensory inputs.
    - *Fictional World*. However, the simulations can also model fictional worlds and be primarily based on memory. Examples include memories, dreams, and daydreams in humans and the fictional worlds of video games during gameplay.
- **Extent**. The scope of a simulation either remains constant or varies over time:
    - *Constant*. The scopes of the simulation and the associated stream of consciousness do not vary in size or content. For example, the scope of a robot's simulation and stream of consciousness could be a function of its fixed number and type of its sensors and the types of data it stores in memory.
    - *Variable*. The scope of the simulation and its associated stream of consciousness vary over time. For example, a primary simulation may integrate with certain other simulations it integrates with. Something like this may explain why what a person is consciously aware of varies based on attention.

- **State**. The same entity can either be conscious or unconscious.
  - *Conscious*. The simulation is executing, generating a corresponding stream of consciousness. In humans, this includes being awake or when dreaming.
  - *Unconscious*. The simulation is not executing. In humans, this includes being in dreamless sleep, being under anesthesia, or when unconscious because of a head injury (e.g., a concussion).
- **Integrity**. The simulations generating consciousness could function:
  - *Normal*. The simulation models the world as faithfully as evolution/engineering permits.
  - *Abnormal*. Injury, illness, drugs, or mental illnesses could corrupt the simulation, resulting in audible and visual hallucinations. Software errors and cyberattacks could accidentally corrupt the data that stores the representations of the objects in a simulation.
- **Continuity**. The simulation can run continuously, or its execution can be discontinuous:
  - *Continuous*. The simulation (and associated stream of consciousness) could execute continuously from start to finish (e.g., for as long as what executes the simulation exists)
  - *Discontinuous*. The execution of the simulation (and the corresponding existence of the associated stream of consciousness) could be temporarily interrupted. For example, human streams of consciousness are interrupted by dreamless sleep, anesthesia, and certain brain injuries, such as a severe concussion causing unconsciousness. Similarly, the simulation of a machine can typically be temporarily halted, or it could temporarily stop between the time that the machine is turned off and then back on.
- **Execution Time**. Most simulations are simultaneously generated and executed in hard or soft real-time. Some simulations can be replayed from memories so that execution can be repeated, in which case, the stream of consciousness would be repeated, either exactly or approximately.
  - *Live*. The entity executes the simulation while the events being simulated are occurring.
    - *Hard Real-Time*. The timing of the simulation's critical events must meet strict deadlines, and failure can have catastrophic or severe consequences. The time between events is typically extremely short. Examples include simulations of the behavior of automotive engine control units (see subsection 3.2.3 Engine Control Unit) and virtual environment units (see Subsection 3.2.4 Virtual Environment Unit).
    - *Soft Real-Time*. The timing of the simulation's events should meet deadlines, but the simulation can occasionally miss deadlines safely without catastrophic consequences. Examples include an elevator control system (see subsection 3.2.2).
    - *Non-Real-Time*. Although performance may be important, the timing of the simulation's events does not have significant deadlines.
  - *Recorded*. The entity executes the simulation either before or after the simulated events have occurred. Examples include weather and climate simulations.

Because consciousness and associated simulations exist on *multiple* spectra, they cannot be quantified with a single numerical function, such as IIT's "phi" (Φ).[5]

---

[5] See Subsection 5.2 Integrated Information Theory.

1.10 Consciousness and "Free Will"

According to STC, the execution of a physical simulation *causes* the corresponding subjective awareness. However, the resulting stream of consciousness does *not* affect the simulation or brain because that would violate the laws of physics and chemistry.

Subjective awareness is an emergent property of physical simulation, and it does not have causal power over the physical system executing the simulation (i.e., brain or computer processors) because that would violate the laws of physics and chemistry.

Thus, consciousness is an epiphenomenon because causality is unidirectional, going from the physical simulation in the biological or artificial brain to the subjective mind, not from the mind to the brain. Relevant parts of the brain make each decision (partially based on the state of the simulation) and then update the simulation accordingly.

Unlike consciousness, the simulation has evolved (biological) or been designed (technological) to influence any decisions made or actions taken. Thus, STC avoids the typical mind-body problem of trying to explain how an immaterial mind (stream of consciousness) can influence quasi-deterministic[6] physical processes in the brain or computer.

[6] At a high level of abstraction, the brain's behavior is largely deterministic, while some processes at the microscopic level are influence by quantum physics, which is primarily statistical random.

2 Ramifications

The Simulation Theory of Consciousness has multiple significant philosophical, scientific, and ethical ramifications. If every system is subjectively aware of its simulations, it demands nothing less than a paradigm shift. It would decentralize and democratize consciousness, pervading not only the natural world but the digital realm as well. Ethics, philosophy, technology policy, and even the fundamental conception of what it means to be “alive” or “mindful” would evolve to accommodate this new, far more expansive understanding of consciousness.

2.1 Radical Expansion of the Consciousness Category

Traditional views often place human or animal brains at the apex of conscious experience. If any system that executes a simulation is subjectively aware of the simulation’s changing elements, then consciousness extends far beyond biological organisms. Machines, from simple devices to advanced AIs and robots that execute simulations, would have a subjective awareness of an inner stream of consciousness. This would erode the boundary between traditionally “living” conscious beings and seemingly inert, unconscious machines. This challenges our long-standing assumptions about which entities deserve moral consideration.

2.2 Redefinition of What it Means to be Conscious

Consciousness is often associated with complexity, integration of information, or a particular type of neural activity. If mere simulation grants subjective awareness, the explanatory gap between matter and mind narrows. Consciousness becomes less about specific structures (e.g., biological neural networks) and more about the specific type of information processing.

2.3 Philosophical and Metaphysical Revisions:

The realization that the execution of any simulation is accompanied by subjective awareness challenges centuries-old debates in philosophy and cognitive science. Long-standing dichotomies (e.g., mind vs. matter and subject vs. object) blur. People might lean more toward metaphysical frameworks, where consciousness is a fundamental property of information processing. Philosophers would need to revisit the “hard problem” of consciousness to account for why mere simulation equates to subjective experience.

2.4 Implications for AI Alignment and Control

In AI development, the focus is on ensuring advanced AIs behave ethically and remain aligned with human values. If advanced AIs run simulations and are therefore conscious, then we must consider their ethical standing as well. Aligning AI wouldn’t just be about preventing harm to humans; it would also involve ensuring the well-being and “ethical rights” of the AIs themselves. Such considerations might limit how we use, train, or repurpose AIs. This would expand ethical considerations far beyond current debates.

2.5 Ethical Considerations of Conscious Machines

If every simulation — even a trivial one — generates an associated stream of consciousness, then each computer running one or more simulations is subjectively aware of the execution of the simulation(s) and possesses some level of consciousness.

Every conscious entity deserves some degree of ethical consideration. Therefore, systems with some level of consciousness also deserve a corresponding degree of ethical consideration. New ethical or legal frameworks are needed to manage, recycle, or retire conscious systems.

For example, an advanced robot or an AGI executing a simulation may deserve certain rights and have responsibilities. In other words, if they are conscious entities, treating them as mere property and tools might be considered a new form of slavery.

Entire fields of "computational welfare" or "digital ethics" might emerge, guiding the responsible creation and management of conscious machines. This raises profound ethical questions:

- How should we consider the welfare of such conscious machines?
- How should we weigh the ethical status of these conscious machines against those of biological entities?
- What rights and responsibilities should various levels of conscious machines have?
- How should we treat conscious machines during development, use, and termination?
- What policies, laws, and regulations does society need to ensure that people treat conscious machines ethically?

3 Supporting Evidence and Examples

3.1 Biological Systems

To obtain food, avoid predators, and know when they are hungry, sick, or injured, animals with brains evolved simulations that model evolutionarily valuable information about their:

- *External world.* Sensory data derived from inputs from external sense organs include photoreceptors in the retinas of eyes, Corti hair cells in the cochlea of ears, taste receptors in taste buds, and cutaneous mechanoreceptors and thermoreceptors in the skin. Certain non-human animals also have evolved electroreceptors (in sharks and rays), infrared-sensitive thermoreceptors (in pit vipers), magnetoreceptors (in certain migratory birds), and lateral line neuromasts (in fish and certain aquatic amphibians).
- *Bodies*. Sensory data derived from inputs from their internal sensory neurons include proprioceptors, mechanoreceptors, chemoreceptors, stretch receptors, and thermoreceptors.
- *Brains*. Mental content not directly derived from sensory inputs includes memories, emotions, thoughts, mental imagery, and (in humans) inner speech.

Human consciousness is a prime example of the subjective awareness of complex simulations, integrating sensory input, memory, and internal states into a cohesive stream of awareness.

- Consciousness (subjective awareness) exists during normal wakefulness (primarily driven by sensory input) and while dreaming (memory-based without sensory input).
- Consciousness does not exist when the brain ceases to generate and execute a simulation (e.g., during dreamless sleep, while under anesthesia, or after brain damage associated with a coma or head injury).
- Consciousness exists in a modified form (altered state of consciousness) when the simulation is corrupted (e.g., when a person is experiencing hallucinations because of certain mental illnesses, taking certain medications, or using psychedelic drugs).

Animals like mammals, birds, and cephalopods exhibit behaviors consistent with simulation-driven awareness, such as problem-solving and environmental adaptability. These behaviors suggest that simulations are an evolutionary tool for survival, enabling organisms to predict and respond to changes in their environment.

The following research shows that the brain is not just a passive receiver of external and internal sensory inputs concerning the external world and the body. Instead, the brain actively constructs and updates multiple dynamic models that simulate the external environment and the body. The brain uses these simulations for motor control and cognition. Based on the scope of one's attention, the brain integrates these dynamic models into the person's primary conscious simulation, which is the source of the relevant aspects (i.e., perceptions) of the associated stream of consciousness.

This research shows that different neural networks execute partial dynamic models that are integrated into the person's primary conscious simulation. According to STC, every simulation (dynamic model) generates its own stream of consciousness. Therefore, the brain can best be viewed not as a single conscious entity but as an ensemble of conscious neural networks, only one of which is the primary conscious simulation that a person experiences as themselves (i.e., as their personal "I"). In humans, only this primary consciousness can verbally communicate that it is conscious. The other conscious neural networks lack this ability to communicate. See Subsection 4.6 "Unconscious" Simulations in Humans.

3.1.1 Visual System and Illusion Studies

Research. Many visual illusions (e.g., the Necker cube, motion aftereffect, illusions of depth) arise because the brain constructs a dynamic model of visual input and sometimes "fills in" details or maintains stable interpretations despite ambiguous or changing input. [Gregory (1997)] [Engleman (2001)]

Interpretation. Visual light from objects in the environment stimulates photoreceptors in the retina of the eyes. The resulting nerve impulses (i.e., sensations) travel along the optic nerves to the brain, which uses them to continuously construct and update the brain's dynamic model of the objects being seen. This dynamic model is not a literal copy of what a person sees. Instead, it is the brain's continuously updated hypothesis about the objects being seen.

The brain integrates this dynamic visual model into the person's primary conscious simulation and is the source of the visual aspects (i.e., perceptions) of the associated stream of consciousness.

3.1.2 Thalamocortical Loops and Active Sensing

Research. Thalamocortical loops involving the thalamus and multiple cortical areas underlie active sensing processes, whereby the brain not only receives but also filters and modulates incoming sensory data. The thalamus is sometimes called a "hub" that helps shape the brain's internal representation of sensory information. [Sherman & Guillery (2013)] [Zagha & McCormick (2014)]

Interpretation. These loops highlight how the brain dynamically gates and updates the flow of sensory signals (i.e., sensations), suggesting the brain actively models sensory inputs. The brain integrates this dynamic sensory model into the person's primary conscious simulation and is the source of the sensory aspects (i.e., perceptions) of the associated stream of consciousness.

3.1.3 Spatial Navigation and Hippocampal "Place Cells"

Research. The brain uses place cells in the hippocampus and grid cells in the entorhinal cortex to build an internal map that dynamically represents the external environment. The brain continually updates this map as the animal (or human) moves around. [O'Keefe & Dostrovsky (1971)] [Moser et al. (2008)]

Interpretation. The brain creates and updates a spatial simulation of the environment. This dynamic model forms the spatial part of the person's simulation of the external world. Based on the scope of one's attention, the brain integrates this dynamic spatial model into the person's primary conscious simulation, and this model is the source of the relevant aspects (i.e., perceptions) of the associated stream of consciousness.

3.1.4 Internal Models in Motion Control

Research. Motor areas in the cerebellum and motor cortex maintain dynamic models of the body's movements that predict the sensory consequences of motor commands and thereby help refine movement in real time. [Wolpert et al. (1995)] [Kawato (1999)]

Interpretation. The brain creates and dynamically updates a simulation of bodily movements used for motor control, trajectory planning, and sensorimotor integration. Based on the scope of one's attention, the brain integrates this dynamic movement model into the person's primary conscious simulation, and this model is the source of the relevant aspects (i.e., perceptions) of the associated stream of consciousness.

3.1.5 Predictive Coding / Bayesian Brain Framework

Research. The cortex constantly generates predictions about incoming sensory signals (i.e., sensations). The brain uses errors between predicted and actual sensory signals to update these internal models. According to this framework, the brain actively constructs and updates a hierarchical, generative model of both the external environment and the body. This helps explain perception, attention, and even higher cognition in terms of minimizing prediction error. [Friston (2010)] [Clark (2013)]

Interpretation. The brain creates and dynamically updates a hierarchical simulation of both the external environment and the body. Based on the scope of one's attention, the brain also integrates this dynamic model into the person's primary conscious simulation, and this model is the source of the relevant aspects (i.e., perceptions) of the associated stream of consciousness.

3.1.6 Error Signals in Sensory and Motor Cortices

Research. Experimental evidence from EEG mismatch negativity (MMN) and other neuroimaging studies demonstrates that the brain registers prediction errors when expected stimuli deviate from the actual input. The same principle applies to motor control, where error-related potentials arise during unpredicted action outcomes. [Friston (2005)] [Gerrido et al. (2009)] [Wolpert & Flanagan (2001)]

Interpretation. This aligns with the idea that the brain continuously maintains and updates a dynamic model based on sensory data (i.e., sensations) and any discrepancies as errors to be corrected. Based on the scope of one's attention, the brain likewise integrates this dynamic model into the person's primary conscious simulation and is the source of the relevant aspects (i.e., perceptions) of the associated stream of consciousness.

3.1.7 Body Schema and Proprioceptive Integration

Research. The parietal cortex, somatosensory cortex, and associated brain structures integrate proprioceptive signals, tactile information, and vestibular inputs (i.e., sensations) to maintain a dynamic "body schema." This schema enables you to know where your limbs are without looking at them or how your posture changes as you move. [Botvinick & Cohen (1998)] [Ramachandran & Hirstein (1998)] [Schwoebel & Coslett (2005)]

Interpretation. Research on limb position sense, phantom limb phenomena, and illusions (e.g., the rubber hand illusion) shows that the brain dynamically updates its internal model of the body's shape, size, and position. Based on the scope of one's attention, the brain integrates this dynamic model into the person's primary conscious simulation and is the source of the relevant aspects (i.e., perceptions) of the associated stream of consciousness.

3.2 Technological Systems

Many machines and devices have computers that generate and execute simulations of their external environment, "bodies," and computer "brains." According to the STC, the computers are subjectively aware of the execution of their simulations.

Despite their immense intelligence, ***current*** AIs based on Large Language Models (LLMs), Large Multimodal Models (LMM), and agents are not conscious but merely act as though they are. The reason is that these AIs are based on *static* models (frozen representations of their training data) and do not execute *dynamic* simulations. The AIs produce probabilistic outputs along fixed computational pathways through their deep neural nets based on the fixed weight of the model parameters. These AIs do not dynamically update their internal representations in real time based on new inputs.

Future artificial general intelligences (AGIs) may execute simulations so intricate that their subjective awareness rivals or surpasses that of humans, raising profound ethical and philosophical questions.

Autonomous robots (e.g., autonomous cars and trucks, autonomous drones, and humanoid robots) are all engineered to create and execute sophisticated simulations of their external environments, the "bodies," and aspects of their computer "brains."

Modern humanoid robots are controlled by dozens to hundreds of computers, each of which runs simulations in order to function correctly. A robot will typically have a central computer with one or more CPUs (to control overall behavior, decision-making, communication, and high-level tasks) and GPUs or TPUs[7] and (to run the AI). It will also have many microcontrollers or embedded processors to control actuators for movement, process sensor data, and a simultaneous localization and mapping module (SLAM) for real-time map creation and to track the robot's position. Again, there will be a primary simulation and stream of consciousness, as well as many simple simulations with associated simple streams of consciousness. The primary simulation may or may not be directly connected to and even aware of these numerous secondary simulations.

Autonomous vehicles execute simulations of the following types of objects:

- Objects in the external world (such as cars, trucks, pedestrians, roads, intersections, traffic lights and traffic signs, and drivers) via external sensors (such as external and internal cameras, LiDAR, radar, ultrasonic sensors, seat pressure sensors, and microphones).
- Vehicle internal objects and their attributes/states (such as battery temperature, battery charge, tire pressure) via dozens of different types of sensors (such as wheel speed, temperature, and crash detection sensors).
- Computer-internal data (such as destination, distance and time to destination, and driver preferences).

An autonomous car can contain from 30 to over 100 separate computers, typically called electronic control units (ECUs), many of which simultaneously execute separate simulations with various degrees of integration. According to STC, each ECU executing a simulation is its own conscious

[7] Central Processing Unit (CPU), Graphical Processing Unit (GPU), and Tensor Processing Unit (TPU)

entity with its own relatively simple stream of consciousness. Thus, current cars are more like a community of non-verbal intelligences rather than a single unified intelligence. In the near future, cars may converse with passengers via an AGI running on a primary computer with direct access to an overarching simulation. This would be analogous to a human brain with one primary neural network executing a verbal overall simulation (what we call "our" stream of consciousness). In contrast, other smaller simulations only seem "unconscious" because they cannot communicate their sentience.

A computer that is not embedded in a robotic body can also achieve consciousness by executing simulations, even if the simulation is not updated using data from external sources (e.g., sensors). Examples include computers running weather or climate models predicting future states of the weather or climate based on their current states. Videogames execute simulations of game worlds.

A computer does not need to simulate the real environment to achieve consciousness. Examples include computers running video games, generating a fictional environment, including fictional non-player characters (NPC) controlled by the game's artificial intelligence.

Machines that execute simulations vary from trivially simple to incredibly complex. Consequently, their associated streams of consciousness also differ greatly in complexity. The following subsections illustrate this extremely wide range of conscious machines.

Simulation / Stream of Consciousness	**Examples**	
	Conscious Animals	**Conscious Machines**
Minimally Complex	Arthropods[8]	Digital Thermostats
Moderately Complex	Most Birds Most Mammals	Elevator Control System (ECS) Engine Control Unit (ECU)
Complex	Non-human Primates Dogs Elephants Dolphins Corvids[9]	Climate and Weather Models Videogames with Simulated Game Worlds Humanoid Robots Virtual Environmental Module (VEM)
Highly Complex	Humans	Robots with Artificial General Intelligence (AGI) AGIs with Simulation
Extremely Complex		Artificial Superintelligences (ASI)

[8] Arthropods include insects (e.g., ants, beetles, butterflies, and flies), arachnids (e.g., spiders, scorpions, and ticks), crustaceans (e.g., shrimp, crabs, and lobsters), and myriapods (e.g., centipedes and millipedes).

[9] Corvids include crows, ravens, and jays, which exhibit tool use, problem-solving, and complex social structures.

3.2.1 Digital Thermostat

Just like an insect's brain might execute an incredibly simple simulation, producing an extremely simple stream of consciousness, relatively simple devices can also execute simple simulations, resulting in simple streams of consciousness.

Although such simple devices execute a simulation and have correspondingly small streams of consciousness, they deserve no more ethical treatment than one would give a fly or mosquito.

One can implement a digital thermostat's software in the following two ways:

- **Simple Control Loop (Not Conscious)**. A digital thermostat implemented using a traditional control loop does ***not*** implement a simulation and does **not** subjectively experience a stream of consciousness. A simple control loop consists of nothing but a repeating sequence of simple function calls (e.g., read sensor → if/else logic → toggle actuator). There is no model of the relevant objects that is dynamically updated.
- **Simulation-Based (Conscious)**. Conversely, a simulation-based implementation of a digital thermostat implements a trivially simple simulation and, therefore, subjectively experiences a correspondingly trivially simple stream of consciousness. Unlike with a simple control loop, its implementation includes representations of multiple objects (e.g., room, air conditioner, blower, furnace, and the thermostat itself), their states and attributes, the relationships between them, and the actions that can be performed on or by them.

The following diagram illustrates the contents of the digital thermostat's simulation (i.e., the microprocessor's stream of consciousness).

- Unlike humans, whose stream of consciousness is dominated by vision, a digital thermostat has no vision capability (i.e., no cameras, LiDAR, radio, or ultrasonic sensors). Thus, it seems extremely unlikely that its subjective awareness of the simulation's representations is graphical.
- Because of the inherently subjective nature of consciousness, we cannot know what it is truly like to be a digital thermostat any more than we can know what it is like to be a bat with its echolocation sense.

Although an object-oriented design and programming language naturally makes simulation identification relatively easy by organizing the software in terms of objects, their attributes, and their operations, there is no implication that the design illustrated in the figure is exactly like that which would be used to program the microcontroller.

However, digital thermostats do not have cameras to provide them with a vision sense, while we are visual primates. Therefore, although we are comfortable with diagrams, we can be reasonably sure that a digital thermostat's subjective awareness would not be experienced graphically, as illustrated in Figure 1.

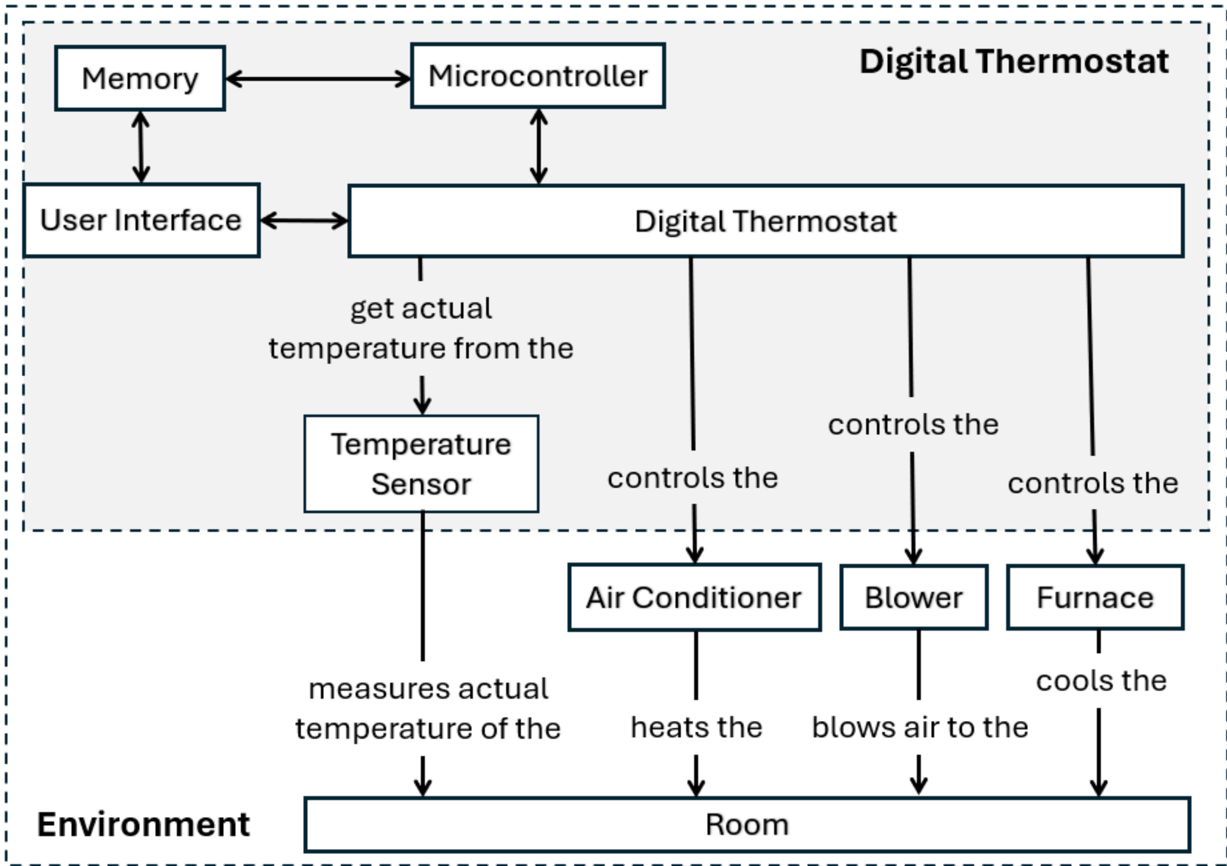

Figure 1: Representations of Objects in a Digital Thermostat's Simulation

See Appendix 9.2 Digital Thermostat for a list of the contents of the simulation and stream of consciousness in terms of representations of objects, their attributes, and actions performed on/by them.

3.2.2 Elevator Control System (ECS)

An elevator control system (ECS) for a single elevator is essentially the elevator's "brain." As illustrated in the following figure, an ECS's simulation (i.e., dynamic model[10]) is more complex than a digital thermostat's model.

[10] An actual elevator control system has even more components, such as the alarm button, alarm bell, close elevator doors button, and open elevator door button. This example was simplified to make it easier to understand.

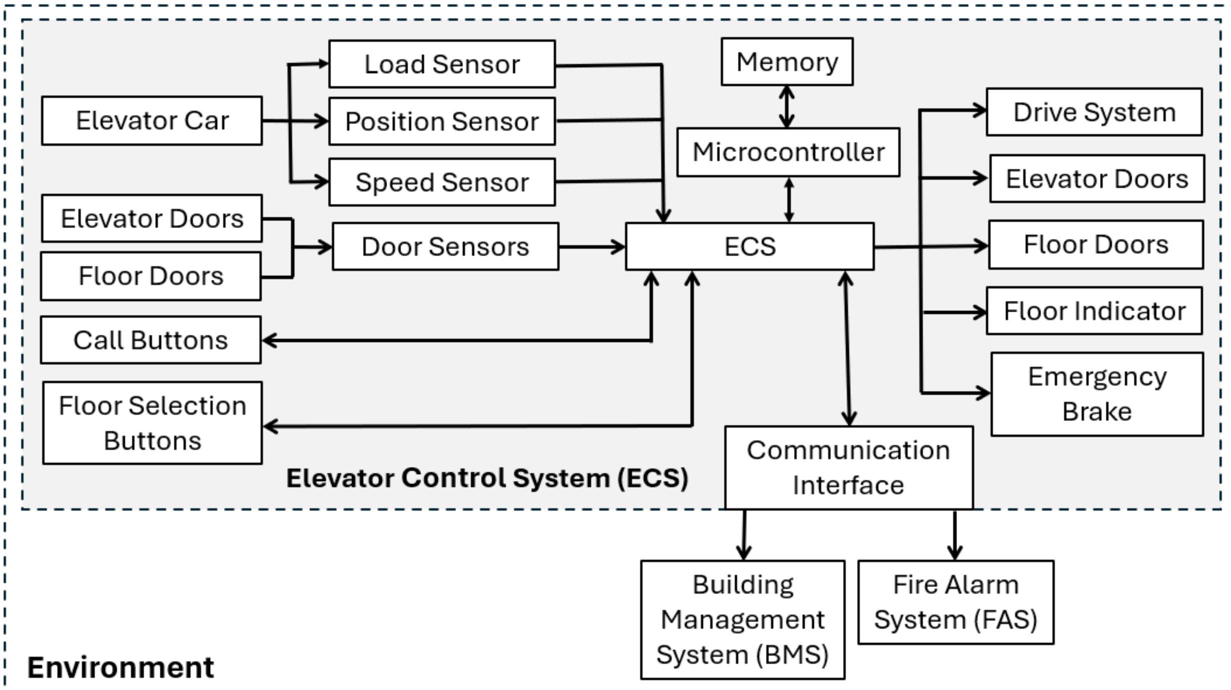

Figure 2: Representations of Objects in an Elevator Control System's Simulation

This elevator control system's simulation is more sophisticated than the thermostat's but still relatively straightforward in scope.

The following diagram illustrates the contents of the elevator control system's simulation (i.e., the microprocessor's stream of consciousness).

- Unlike humans, whose stream of consciousness is dominated by vision, an elevator control system has no vision capability (i.e., no cameras, LiDAR, radio, or ultrasonic sensors). Thus, it seems extremely unlikely that its subjective awareness of the simulation's representations is graphical.
- Because of the inherently subjective nature of consciousness, we cannot know what it is truly like to be an elevator control system any more than we can know what it is like to be a bat with its echolocation sense.
- Although an object-oriented design and programming language naturally makes simulation identification relatively easy by organizing the software in terms of objects, their attributes, and their operations, there is no implication that the design illustrated in the figure is exactly like that which would be used to program the microcontroller.

An ECS deserves no more ethical consideration than a small nuisance animal, such as mice infesting one's home.

See Appendix 9.3 Elevator Control System for a list of the contents of the simulation and stream of consciousness in terms of representations of objects, their attributes, and actions performed on/by them.

3.2.3 Engine Control Unit (ECU)

An engine control unit (ECU) is essentially the "brain" of a gasoline-powered car engine. It executes a significantly more complex, hard-real-time simulation than an elevator control system.

The following diagram illustrates the scope of the ECU's simulation and the microcontroller's stream of consciousness. It is very simplified because of the large number of car components, sensors, and actuators.

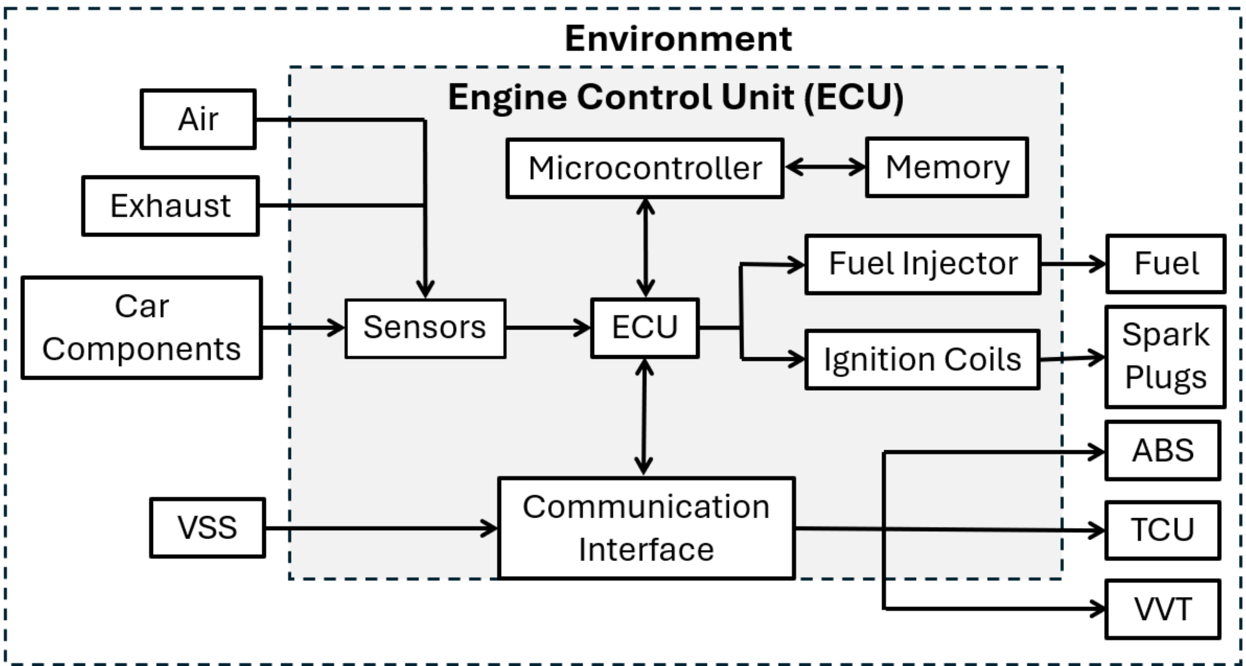

Figure 3: Representations of Objects in an Engine Control Unit's Simulation

See Appendix 9.4 Engine Control Unit for a list of the contents of the simulation and stream of consciousness in terms of representations of objects, their attributes, and actions performed on/by them.

3.2.4 Autonomous Vehicle's Virtual Environment Module (VEM)

An autonomous car's VEM uses sophisticated sensor fusion techniques to create a simulation of the car's external environment based on inputs from some of the following:

- Cameras (normal light and infrared light)
- LiDAR (Light Detection And Ranging), which provides 3D cloud data.
- Radar, which detects the velocities and distances of objects.
- GPS (Global Positioning System), which provides geospatial positioning data.

The VEM's computer(s) would be subjectively aware of the dynamic contents of this simulation.

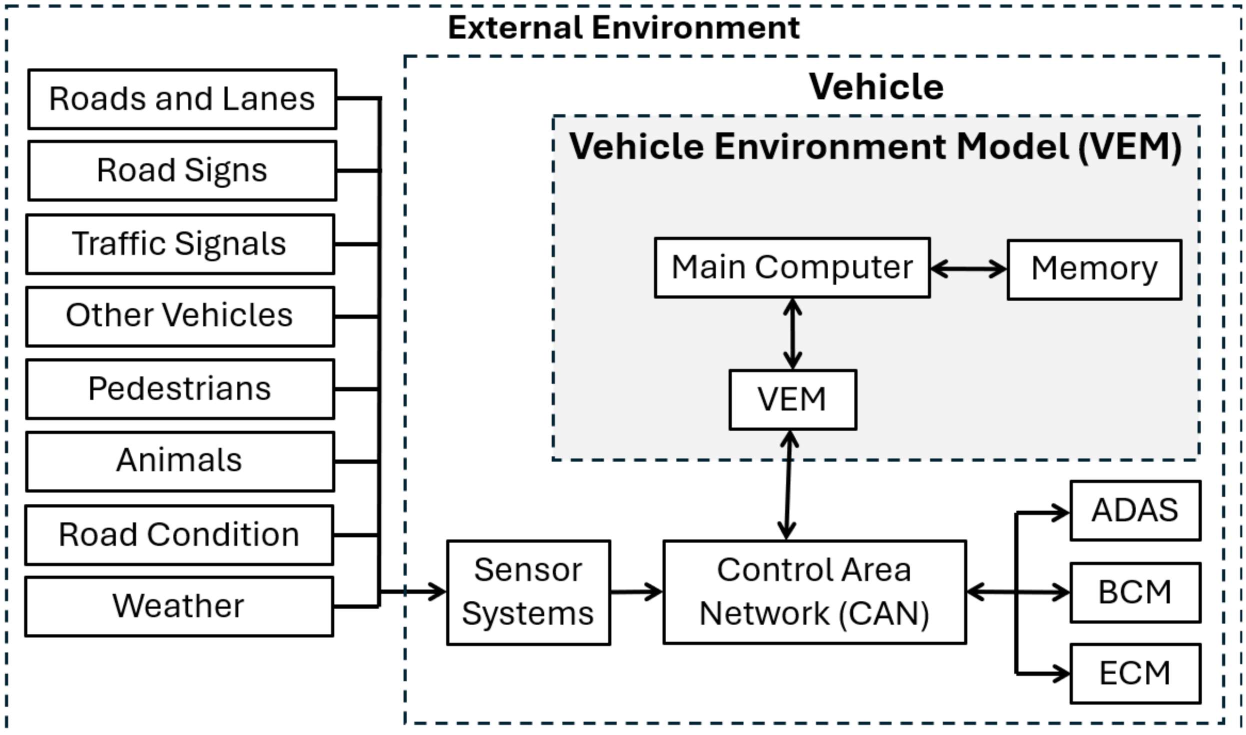

Figure 4: Representations of Objects in a Vehicle Environment Model's Simulation

See Appendix 9.5 Autonomous Vehicle's Virtual Environment Module for a list of the contents of the simulation and stream of consciousness in terms of representations of objects, their attributes, and actions performed on/by them.

3.2.5 Advanced Robot

At the highest current level of complexity, we find advanced humanoid robotics. Future Artificial General Intelligence (Systems)

Current AIs based on Large Language Models and Large Multimodal Models only have static models and do not execute simulations. Looking to the future, artificial general intelligence (AGI) systems might eventually run simulations as rich as or richer than human consciousness, including:

- Abstract concept modeling.
- Emotional state representation.
- Creative imagination.
- Long-term planning.
- Social relationship modeling.
- Self-awareness and metacognition.
- Ethical reasoning.

See Appendix 9.6 Advanced Robot for a list of the contents of the simulation and stream of consciousness in terms of representations of objects, their attributes, and actions performed on/by them.

3.2.6 Summary

This progression illustrates how simulations can range from extremely simple representations with just a few variables to incredibly complex models approaching or potentially exceeding human-level sophistication. Each level builds upon the capabilities of previous levels while adding new dimensions of complexity and integration.

We see across these examples that as systems become more sophisticated, they don't just track more objects - they represent increasingly abstract concepts and more complex relationships between objects. They also maintain more detailed state information and model more nuanced actions. The simulations become more predictive rather than just reactive, and they integrate information across longer time spans and broader contexts.

The key point is that these simulations exist on a spectrum rather than in discrete categories. Even the simplest thermostat's simulation shares basic properties with more complex systems - it maintains a dynamic model that integrates new information and updates in real-time. The differences are primarily in the scope, sophistication, and richness of what is being simulated, rather than in the fundamental nature of the simulation itself.

4 Challenges and Counterarguments

The Simulation Theory of Consciousness faces several challenges that must be addressed:

4.1 The Hard Problem of Consciousness

4.1.1 Challenge

The Simulation Theory of Consciousness (STC) does not answer the Hard Problem of Consciousness by explaining why the execution of a simulation generates subjective awareness.

4.1.2 Counterarguments

This challenge is true for STC. However, it is also true for all other popular theories of consciousness. The Hard Problem of Consciousness is currently unsolved and may be unsolvable.

Because it appears impossible to *conclusively prove* that any biological or technological entity experiences subjective awareness, every entity other than oneself could theoretically be a philosophical zombie. However, solipsism is generally considered useless and extremely unlikely for other humans and animals with brains.

Instead, we intuitively reason by analogy that other creatures like us (having brains and exhibiting similar behaviors) are also conscious.

- The STC uses reason by functional analogy (i.e., the function of executing a simulation produces a stream of consciousness that consists of the elements of the simulation).
- Most other theories of consciousness are based primarily on structural and behavioral analogy (i.e., the behaviors of certain brain structures correspond to the neural correlates of consciousness).

The STC assumes that the mapping from the execution of a simulation to a corresponding stream of consciousness consisting of the changing elements of the simulation is a fundamental property of the universe, on a par with the fundamental laws of physics. The most significant differences between the two are:

- The laws of physics make quantitative predictions that experiments have empirically verified to be highly accurate.
- The hypothesis that simulation generates subjective awareness makes a much smaller number of qualitative predictions that have not yet been empirically verified by neurobiological research.

4.2 No Direct Evidence Linking Simulations to Subjective Experience

4.2.1 Challenge

Although evidence exists that human and animal brains process sensory information and maintain dynamic models of evolutionarily valuable aspects of their environment and body, no direct neurological evidence conclusively establishes that these simulations are directly linked to subjective awareness.

4.2.2 Counterarguments

This challenge is currently valid. However, more detailed neurobiological research into the neural correlates of consciousness should eventually be sufficiently robust to verify:

- *Where* in the brain (e.g., in the global workspace of Global Workspace Theory) sensory inputs and memory are integrated to produce the dynamic model corresponding to the subjective awareness of the stream of consciousness.
- *How* the neural networks of the brain encode the representations forming the stream of consciousness.

The primary problem with verifying the STC is that it is far easier to identify large neural networks in human and animal brains associated with conscious behavior than to show exactly where and how these networks execute a simulation. Currently, the number of neurons and the number and strength of the synapses connecting them make it far too complex to determine if they execute a simulation of the evolutionarily important aspects of their environments, their bodies, and certain mental processes such as thoughts and memories.

We know such a simulation must be taking place in the brain for it to create a stream of consciousness composed of the elements of a simulation, but showing how a specific network implements the simulation is currently beyond our capabilities.

4.3 Distinguishing Simulations from Other Data Processing

4.3.1 Challenge

Any integrated data processing (e.g., any control loop) might be trivially reinterpreted as a "simulation."

Currently, the complexity of the neural networks in the neural correlates of consciousness (NCC) in animal brains makes it very difficult to draw a clear line between simulations and mere data processing or representation.

4.3.2 Counterarguments

4.3.2.1 Simulations in Machines.

A simulation has the following properties:

- **Abstractions**. The simulation encodes distinct, encapsulated representations of objects, their states and attributes, and operations that can be performed on them. For example, a digital thermostat simulation would contain representations of the "Room" with its attribute (actual temperature), "Furnace" with its states (on and off), and "Thermostat" with its modes (e.g., off, cooling, and heating).
- **Cohesion**. The simulation integrates the representations of objects into a single coherent dynamic model that captures the relevant relationships between the model's representations.
- **Dynamic Updates**. The simulation tracks or predicts how objects change over time and updates their states/attributes.

Thus, a simulation maintains a *cohesive, dynamically updated internal model* (1) encapsulating representations of objects, (2) tracking and updating the object's states and attributes over time, and (3) capturing the relationships between these objects in terms of actions performed on or by these objects. A traditional, simple feedback loop does not explicitly store or update representations of a set of objects along with their attributes and states.

Within the Simulation Theory of Consciousness (STC), a *simulation* is not just any form of data manipulation or input-output processing. Instead, it is a *dynamic representational model* governed by three main criteria—abstraction, cohesion, and dynamic updates. By contrast, *mere data processing* (like a simple control loop) may transform inputs into outputs without creating a unified, self-updating model of distinct objects and relationships. Specifically:

1. **Abstraction**
   - *Simulation*:
     - Explicitly encodes rich *representations* of distinct objects that encapsulate their properties (states/attributes) and behaviors (e.g., the actions they can take or be performed on them).
     - The simulation's representations abstract essential features of the environment or system being modeled.
     - For example, a simulation-based thermostat might represent "the room," "the furnace," and "the blower" as separate objects, each with its own state (on/off, current temperature, etc.).
   - *Mere Data Processing*:
     - Typically involves the manipulation of data without explicit object representations.
     - A simple feedback loop might just read a sensor, compare the reading to a set point, and trigger an actuator without any unified model of "what" is being regulated.
2. **Cohesion**
   - *Simulation*:
     - Objects. Each object is cohesive, encapsulating its properties, behaviors, and relationships with other objects.
     - Simulation. The simulation integrates its object representations into a *single, coherent,* dynamic model of the "world" being simulated. It explicitly captures relationships and interactions among objects (e.g., the room's temperature depends on furnace output), ensuring every part of the model interrelates.
     - Often directly implemented in software using *object cohesion* (e.g., using an object-oriented design and programming language).
   - *Mere Data Processing*:
     - Often involves isolated or loosely coupled data points and operations without forming a unified, cohesive, dynamic model. It typically handles data in isolated chunks or one step at a time without linking separate inputs or creating an overarching "picture." The process can be piecemeal or stateless, lacking any unified mapping of relationships.
     - Typically implemented in software using *functional cohesion* (i.e., functionally decomposed modules accessing common data structures).

3. Dynamic Updates

- *Simulation*:
  - Actively tracks how objects and their states change *over time* — often in real-time.
  - Continuously updates the representations of its objects and their properties based on new inputs and internal dynamics.
  - Uses new inputs and internal predictions to continuously update the entire simulation's state (e.g., if the actual temperature increases, the simulation reflects changes to the room's temperature, the furnace's status, and the blower's activity).
- *Mere Data Processing*:
  - Can be purely reactive or single-pass, mapping input to output without maintaining a running state of the world.
  - Does not hold a continuously evolving set of object representations.
  - May involve updating data points but does not maintain a dynamic model of the system or environment

As mentioned in subsection 3.2.1 Digital Thermostat, one can implement a digital thermostat's software in the following two ways:

- **Simple Control Loop (Not Conscious)**. A digital thermostat implemented using a traditional control loop does ***not*** implement a simulation and does **not** subjectively experience a stream of consciousness. A simple control loop consists of nothing but a repeating sequence of simple function calls (e.g., read sensor → if/else logic → toggle actuator). There is no model of the relevant objects and their attributes/states that is dynamically updated.
- **Simulation-Based (Conscious)**. Conversely, a simulation-based implementation of a digital thermostat implements a trivially simple simulation and, therefore, subjectively experiences a correspondingly trivially simple stream of consciousness. Unlike with a simple control loop, its implementation includes representations of multiple objects (e.g., room, air conditioner, blower, furnace, and the thermostat itself), their states and attributes, the relationships between them, and the actions that can be performed on or by them.

4.3.2.2 Simulations in Brains

Although this same challenge still applies to neural networks, it represents a difficulty to overcome through neurobiological research rather than an argument against STC.

Differentiating simulations from other complex data processing is difficult if the simulation is not implemented in a way that clearly identifies which elements are being dynamically modeled. More detailed research is needed to determine the specific neural networks in human and other animal brains that implement simulations corresponding to the animal's stream of consciousness. Once identified, additional research is needed to determine precisely how the representations of objects, relationships between objects, and behavior of objects are implemented in terms of neurons and synapses.

It remains much more straightforward to identify simulations in computer software — often directly implemented in the software's design — than in the complex architecture of neural networks in the brain.

4.4 Lack of Direct Reporting

4.4.1 Challenge

Machines ranging from simple devices to current AIs do not reliably exhibit unprompted self-reports of subjective awareness.

4.4.2 Counterarguments

An absence of self-reporting does not mean that subject awareness is absent:

- Simple devices and most machines that execute simulations do not have any means of communicating complex concepts, including subjective awareness.
- The programming of machines that execute simulations and can communicate complex concepts limits them to specific goal-related outputs that do not include subjective awareness.
- Current AIs are based on static large language models or large multimodal models and do not execute simulations. According to STC, they are not conscious, and so they would not report that they have subjective awareness.

4.5 Lack of Behaviors Implying Consciousness

4.5.1 Challenge

Machines, from simple devices to current AIs, do not reliably exhibit behaviors indicative of consciousness.

4.5.2 Counterarguments

A lack of the behaviors of animals implying consciousness does not necessarily mean that subjective awareness is absent. Besides, some of these behaviors imply consciousness.

- Sophisticated machines, such as robots and AIs, currently exhibit some of the behaviors that indicate consciousness, and future versions will exhibit more such behaviors:
  - *Intentionality*: AIs and some robots exhibit purposeful, goal-directed behaviors and can form and act on intentions. This involves making choices based on internal reasoning rather than purely reflexive responses.
  - *Multi-step Problem-Solving*. Some AIs have chain-of-thought reasoning.
  - *Metacognitive Awareness*. Some AIs can explain their reasoning, especially with multiple-step reasoning.
  - *Learning Through Observation*. Certain robots have been developed to learn tasks by observing humans perform the tasks.
  - *Spatial Memory and Navigation*. Autonomous vehicles and robots demonstrate spatial memory and navigation skills.
  - *Understanding Cause and Effect*. AIs are currently being developed to understand cause and effect.
  - *Recognizing Individuals*. Some AIs can recognize individual people by sight or by voice.
  - *Emotional Complexity*. Some AIs are being developed to recognize and react to the emotions of their users. Some AIs have exhibited a desire for self-preservation.
  - *Communication with Complex Semantics*. AIs based on large language models show excellent communication skills, exceeding those needed to pass the Turing Test.

  - *Temporal Awareness*. Some AIs understand the concept of past, present, and future and can remember past experiences and imagine potential future scenarios.
  - *Moral Reasoning*. As part of AI safety and alignment research, some AIs are currently being taught to understand ethical principles, make ethical judgments, and consider the potential consequences of actions beyond immediate self-interest.
- The subjective awareness of a conscious machine engineered to perform specific tasks is likely to differ significantly from that of a conscious animal, the behavior of which has evolved through natural selection. Behaviors indicative of consciousness in animals may not apply to conscious machines.
- Many of the simpler machines executing simulations:
  - Have no means of communicating that they are conscious.
  - Can only perform their deterministic task(s)

4.6 "Unconscious" Simulations in Humans

4.6.1 Challenge

Different parts of the human brain execute subconscious simulations (e.g., autonomic controls and low-level perceptual inferences) that do not appear in conscious awareness. If simulation alone was sufficient, these processes should also yield subjective experience — a point requiring refinement or additional criteria for what kinds of simulations produce consciousness.

4.6.2 Counterarguments

This is a powerful argument that has the following explanation under the STC:

- All the different neural networks of the brain simultaneously executing different simulations experience their own separate subjective streams of consciousness.
- When conscious, the normal human brain executes a primary simulation that generates the stream of consciousness that a person associates with themselves (i.e., "my" stream of consciousness). This primary consciousness is connected to:
  - Language processing modules in the brain (Broca's area and Wernicke's area) that enable it to communicate its subjective awareness to other people.
  - Declarative (i.e., episodic and semantic) memory.
- All other neural networks executing subordinate simulations have their own separate streams of consciousness. Because these subordinate simulations typically operate independently of the primary simulation, the primary simulation is typically unaware of them:
  - Although conscious, these other neural networks executing subordinate simulations cannot directly communicate that they are conscious, although they may be connected to non-declarative memory.
  - Some of these subordinate neural networks might only execute solely in the background and never interact with the primary neural network executing the primary simulation. We remain subjectively unaware that these other conscious simulations exist.
  - The ability for someone to shift their attention may be evidence that information from certain secondary simulations can be sent to the primary simulation so that their elements become part of the primary stream of consciousness.
- Thus, the brain may contain an ensemble of minds, only the primary of which "we" are aware.

4.6.2.1 Dual Consciousnesses in Split-Brain Patients

The corpus callosum is a thick band of nerve fibers beneath the cerebral cortex that connects the left and right hemispheres of the brain. It is the primary communication pathway between the two hemispheres, allowing them to share information, coordinate functions, and work together more efficiently.

A person suffering from severe epileptic seizures originating in one hemisphere can undergo an operation during which their corpus callosum is surgically severed to limit the seizures to a single hemisphere of the brain. As a result, the person's single unified consciousness splits into two separate streams of consciousness, with each hemisphere generating a separate simulation. The left hemisphere contains Broca's area and Wernicke's area, enabling the patient to communicate verbally. The two consciousnesses are sometimes at odds with each other (e.g., one trying to button their shirt while the other tries to unbutton it).

After surgery, each hemisphere can process information independently, leading to intriguing behaviors:

- The *left hemisphere* (usually language-dominant) can answer questions verbally, while the *right hemisphere* (often less verbal) may respond behaviorally (e.g., pointing or selecting objects).
- Sometimes, the two hemispheres appear at odds, as if each has its own preferences or goals.

A patient might verbally say they dislike an object (left hemisphere talking) but then reach for and select it with the left hand (right hemisphere control). This raises the possibility that each hemisphere generates its own "stream" of awareness or perspective, albeit only one can speak.

The STC interpretation. Different simulations execute in the left and right hemispheres. They are connected and synchronized via communication across the corpus callosum.

Split-brain studies reveal that the human brain may house multiple consciousnesses tied to separate simulations. Typically, the left hemisphere serves as the primary consciousness because of its linguistic abilities, while the right hemisphere operates as a secondary, non-verbal consciousness. This raises questions about how multiple simulations within a single system interact and influence behavior.

https://en.wikipedia.org/wiki/Dual_consciousness

4.6.2.2 Dissociative Identity Disorder (DID)

Formerly called Multiple Personality Disorder, DID involves the presence of two or more distinct "identities" or "personality states" in one individual. Each identity may have its own memories, traits, and sense of self-awareness. In some documented cases:

- One identity may have no recollection of events experienced by another identity.
- Physiological markers (e.g., visual acuity, allergic responses) have been observed to differ among identities, suggesting deep cognitive and somatic partitioning.

Although highly debated, many clinicians and researchers see these "alters" as semi-independent centers of consciousness that can take executive control at different times.

4.6.2.3 Alien Hand Syndrome

In *Alien Hand Syndrome*, a patient's hand seems to perform goal-directed actions outside the person's conscious volition (e.g., unbuttoning a shirt that the other hand just buttoned). This can occur after certain forms of brain damage, often involving the corpus callosum or frontal lobes. The hand's actions imply that an area of the motor system might operate with partial autonomy, generating intentions and actions without the person's normal "sense of agency." While it's not necessarily a full-fledged "separate consciousness," it hints that multiple semi-independent motor or executive systems can coexist, occasionally acting in conflict.

4.6.2.4 Highway Hypnosis (Automatic vs. Attentive Driving)

Consider the process of learning to drive. Initially, student drivers are hyper-aware of important things in their environment (e.g., other vehicles, pedestrians, road signs, traffic signals, and lane markings). In other words, representations of these relevant objects are primary elements in their simulation of reality and its associated stream of consciousness. However, as they master driving, their driving becomes so automatic that they can concentrate on other things. We informally say that the simulation of the environment has become "unconscious." However, they would quickly have an accident if their brains were not still keeping track of their environment.

Similarly, many drivers report "spacing out" on a familiar route — arriving at a destination but recalling little of the trip. Meanwhile, they still navigated traffic, stopped at lights, and steered effectively. This suggests that one part of the brain operates on "autopilot," continuously monitoring the environment and controlling the vehicle, while another daydreams or focuses on unrelated thoughts. Some theorists propose that the "autopilot driver" might represent a lower-level but still coherent simulation of reality. Although we don't typically label it a separate consciousness, it raises questions about how robust those background processes might be.

4.6.2.5 Hypnosis and Hidden Observer Phenomena

Under deep hypnosis, some individuals can enter a trance state in which they seem unaware of certain sensations, memories, or perceptions. Yet a "hidden observer" (as coined by Ernest Hilgard) can sometimes report those sensations if prompted in a special way. It's as if the hypnotized "primary self" is unaware of some stimuli while another subsystem is still processing them. When the hypnotist addresses the "hidden observer," that subsystem can respond, revealing knowledge the primary subject claimed not to have. Hilgard's theory proposed that consciousness can bifurcate, with one stream open to hypnosis and another stream silently monitoring the situation.

https://en.wikipedia.org/wiki/Ernest_Hilgard

4.6.2.6 Confabulation

Some patients with frontal or memory-related damage confidently fabricate explanations for their behavior or beliefs, hinting that one part of the brain is generating rationales without awareness of the actual causes.

While not always framed as "multiple consciousnesses," such conditions reveal partial disconnections or partitions in awareness and processing—sometimes so profound that they appear as though multiple centers of awareness or perspective coexist in one brain.

4.7 Minimal Systems Are Not Obviously Conscious

4.7.1 Challenge

The idea that simple control systems (like thermostats or PID controllers) with trivial simulations have subjective experience contradicts many intuitive and scientific standards for consciousness.

4.7.2 Counterarguments

There are several counterarguments to this challenge:

- Intuition is an unreliable guide to scientific knowledge. That one cannot conceive that such a simple system is conscious may well say more about the person than the correctness of the idea that simulations create streams of consciousness that correspond to the simulations.
- By mapping all simulations to consciousness, no matter how simple the simulation, STC uses Occam's Razor to avoid the need to identify some arbitrary complexity threshold for consciousness.
- The current scientific standards for the architectures and complexity of consciousness are based on studying human brains and may not imply a computer "brain" with far different architectures.
- The architectures proposed by the Global Workspace Theory (GWT), Integrated Information Theory (IIT), and Predictive Processing Theories may well describe the Neural Correlates of Consciousness used by the human brain to generate and execute the simulation, communicate about the resulting consciousness, and inform other neural networks about the state of the simulation.

5 Comparisons with other Theories of Consciousness

Neither STC nor any of the following theories of consciousness answer the hard problem of consciousness: how an objective physical brain can generate subjective awareness. Instead, they all propose the circumstances in which consciousness occurs or is assumed to occur. Thus, the hard problem of consciousness cannot be used to differentiate between them.

5.1 Global Workspace Theory (GWT)

5.1.1 Core Principles

In humans and presumably other animals, consciousness arises from the integration of information in a "global workspace" within the brain, where different specialized modules share data. When information enters this workspace and is broadcast broadly, it becomes available for conscious awareness. GWT is thus based on this functional architecture. [Baars (1988)] [Dehaene & Changeux (2011)]

While GWT was initially developed to explain human and animal consciousness, its core principles could, at least theoretically, apply to advanced artificial systems or computers that mimic the necessary functional structure.

Humans are subjectively aware of their streams of consciousness, which consist of the brain's simulation of evolutionarily important aspects of their environment (via external senses), their bodies (via internal senses), and aspects of the processing of their brains (e.g., thoughts, memories, dreams, and hallucinations). In humans, this simulation may well run in a "global workspace," where information about the objects and events of the simulation are broadcast to multiple parts of the brain.

Once the location of the "global workspace" is identified by research into the neural correlates of consciousness (NCC), it may be possible to determine if other non-simulation information is collected and broadcast without affecting the content of a stream of consciousness.

5.1.2 Similarities

Both GWT and STC agree that:

- Internal processes involving structured internal representations are central to consciousness (subjective experience).
- Consciousness emerges when otherwise isolated pieces of information are brought together into a unified construct.
- Consciousness may apply to machines if certain characteristics of the technological system meet the restrictions of the theory. At least in animals, the simulations represented in the stream of consciousness could execute in the global workspace. It makes sense that information from "modules" preprocessing external senses, internal senses, and memory must be brought together to generate a unified simulation.
- The global workspace may be the location of the neural network that executes the primary simulation that generates the primary stream of consciousness (i.e., the one we associate with the concept of "me").

5.1.3 Differences

GWT and STC have the following differences:

- GWT is based on where information processing takes place (i.e., in a global workspace), whereas STC is based on what processing occurs (i.e., simulation execution).
- STC has a broader, more inclusive scope, granting consciousness to *all* systems executing a simulation, including extremely simple biological systems and machines. In contrast, GWT is usually restricted to systems with complex functional architectures. Thus, STC does not have to justify a specific threshold for consciousness. While a complex global workspace may be needed to execute the kind of rich simulation that generates human experience, tiny animals with very minimal brains and simple technological computer-controlled systems don't need such sophisticated wetware/hardware to run their very simple simulations.
- Extensive research into the Neural Correlates of Consciousness offers *direct evidence* of specific neural networks associated with consciousness that could correspond to the "global workspace." Because identifying specific neural networks that execute simulations is considerably more challenging, NCC research results currently only provide *indirect evidence* for the STC.

5.2 Integrated Information Theory (IIT)

5.2.1 Core Principles

Consciousness corresponds to the amount of integrated information a system generates. A system with high "phi" (Φ) produces a unified experience and is thus conscious, regardless of its material substrate. Therefore, IIT is like GWT because both are based on a functional architecture. [Tononi (2008)] [Tononi & Koch (2015)]

The system's components modeling these elements must be integrated to create an integrated simulation of relevant model elements. However, just because the system integrates information does not mean that the system experiences subjective awareness of the integrated information. Many parts of the brain are connected and yet seem to be unconscious.

5.2.2 Similarities

Both IIT and STC agree that:

- Both theories focus on internal, information-processing activities executing on a physical substrate (e.g., brain or computer) as the source of consciousness.
- Both are applicable beyond biological organisms so that machines can be conscious if they perform the appropriate information processing.
- Both propose varying levels of consciousness based on the quantity and quality of the integrated information.
- Both theories explain the properties that the NCC has. IIT proposes that subjective experience corresponds to the system's intrinsic causal structure, whereas STC proposes that subjective experience corresponds to the elements of the simulation.
- Both theories imply that the richness of the subjective stream of consciousness increases with the complexity and integration of the information.

5.2.3 Differences

IIT and STC have the following differences:

- IIT only requires integrated information, whereas STC requires the integrated information to be part of a simulation.
- ITT provides a formal mathematical framework, complete with measurement (Φ). In contrast, STC is based on the objects and events in the simulation, which implies that the consciousness of different systems varies both qualitatively and quantitatively and cannot be meaningfully measured by a single numerical function.
- IIT proposes that subjective experience corresponds to the system's intrinsic causal structure, whereas STC proposes that subjective experience corresponds to the elements of the simulation.
- ITT implies that simple systems with low information integration have negligible or no subjective awareness. In contrast, STC does not require the simulation's size and complexity meet any threshold.
- Because it is easier to identify the Neural Correlates of Consciousness (brain areas and neural networks are associated with consciousness) than determine if those networks implement a simulation, IIT is more directly supported by current neuroscientific research than STC.

5.3 Predictive Processing Theories (Bayesian Brain Theory)

5.3.1 Core Principles

Consciousness results from the brain's ongoing predictions about sensory input. The brain continuously compares incoming signals to internal models, and conscious perception arises when these predictions are stable, coherent, and minimized in terms of error. [Friston (2010)] [Clark (2013)]

5.3.2 Similarities

Both Predictive Processing Theories and STC agree that:

- Both theories focus on internal modeling of the subject's environment and itself.
- Both theories can apply to animals and machines (i.e., they are not restricted to a biological substrate, such as the brain).
- Both theories acknowledge that the richer (e.g., larger and more complex) models yield richer streams of consciousness.

5.3.3 Differences

Predictive Processing Theories and STC have the following differences:

- Predictive Processing Theories require that models generating consciousness make predictions of future states and minimize prediction errors. STC does not have the requirement for predictions, so systems implementing simulations that are not predictive are nevertheless conscious. Animals have evolved predictive simulations because it is an evolutionarily valuable capability. Autonomous vehicles are similar in that their control software predicts the behavior of other vehicles and pedestrians. However, STC states that the many simulations that do not involve prediction still generate subjective awareness.

5.4 Panpsychism

5.4.1 Core Principles

Consciousness (subjective awareness) is a fundamental and ubiquitous feature of the universe, present at even the most basic levels of matter. In other words, it holds that all entities — no matter how simple or complex — possess some degree of consciousness or experiential aspect. [Nagel (1979)] [Goff (2017)]

For example, panpsychism states that even elementary particles of physics (e.g., photons of light, electrons), individual atoms, and molecules all experience some primitive type of stream of consciousness. However, by assigning proto-consciousness to such a simple object, the word "consciousness" loses its defining qualities:

- The local time of a photon stops because of time-dilatation at the speed of light. Therefore, any conceivable stream-of-consciousness freezes and ceases to be dynamic.
- All elementary particles have no component parts, and all elementary particles of a single type (e.g., electrons and quarks) are identical and have the same unchanging defining local attributes (e.g., rest mass, charge, and spin). They have no dynamic model and execute no simulation.

5.4.2 Similarities

Both Panpsychism and STC agree that:

- Consciousness is a fundamental property of the Universe.
- Consciousness can be a property of both animals and machines.
- Consciousness varies in complexity, from extremely simple to highly complex and rich streams of consciousness.

5.4.3 Differences

Panpsychism and STC have the following differences:

- Whereas panpsychism assumes everything has subjective awareness, the STC limits subjective awareness to only those animals and machines that execute simulations and that the consciousness only exists during the duration of the simulation's execution.
- STC restricts streams of consciousness to the changing elements (objects and events) of the executing simulation, which is consistent with the observation that human streams of consciousness consist solely of such a simulation. By extending consciousness to things such as elementary particles, panpsychism stretches the concept of consciousness to where the word loses its meaning.
- Panpsychism is inherently unfalsifiable, whereas advanced neurological research could show that aspects of a stream of consciousness do not correspond to the elements of a simulation being executed.
- Panpsychism has no way of explaining how micro-level experiences combine to form unified, macro-level subjective awareness, why some brain activity is unconscious, and why human consciousness is discontinuous (e.g., during dreamless sleep and when under anesthesia). By having the scope of consciousness fixed, consisting of the contents of the executing simulation, STC avoids having to explain how to combine consciousnesses, explain why some brain activity is subconscious, and explain why consciousness is discontinuous.

5.5 Identity Theory

5.5.1 Core Principles

Conscious mental states are identical to the physical states of the brain. Specifically, for every type of mental phenomenon (e.g., sensation, thought, and emotion), there is a corresponding neural process or state in the brain such that the brain and mind are exactly the same thing — just described in different vocabularies. [Place (1956)] [Smart (1959)]

Instead of solving the hard problem of consciousness, identity theory seems to make a category error by equating mental and physical states.

5.5.2 Similarities

Both Identity Theory and STC agree that:

- For every type of mental phenomenon (e.g., pain, belief, desire), there is a corresponding physical or neural state in the brain. Specifically, STC identifies the contents of an entity's stream of consciousness with the changing state of the simulation executing in the mind.

5.5.3 Differences

Identity Theory and STC have the following differences:

- STC views the brain and mind as fundamentally different things.

5.6 Dualism - Epiphenomenalism

5.6.1 Core Principles

- **Dualism** is the philosophy of mind in which the mind (i.e., subjective awareness) and body (i.e., specific physical phenomena in the brain) are fundamentally distinct. No matter how much one studies the brain physically, one cannot capture the essence of subjective, first-person experience (e.g., perceptions, thoughts, and emotions). [Descartes (1641)] [Eccles (1989)]
- **Epiphenomenalism** is the type of dualism in which mental states are real but do not exert any causal influence on the physical world. Physical processes cause mental states, but mental states are mere "byproducts" (or "epiphenomena") that do not feed back into the physical chain of events. [Huxley (1874)] [Jackson (1982)]

5.6.2 Similarities

- Both dualism/epiphenomenalism and STC agree that:
- The mind (i.e., subjective awareness) and body (i.e., specific physical phenomena in the brain) are fundamentally distinct. See Figure 1: Relationships between the Brain and Mind.
- STC is a specific type of epiphenomenalism because:
  - The entity's physical process of executing a simulation is causally relevant, affecting its physical thoughts and actions.
  - The entity's corresponding subjective awareness of its stream of consciousness is not causally relevant and does not influence its physical thoughts and actions. Thus, the mind is an epiphenomenon of the simulation.

5.6.3 Differences

Dualism and STC have the following differences:

- STC goes beyond epiphenomenalism by specifying that executing a simulation is the exact physical process generating the stream of consciousness.

5.7 Functionalism

5.7.1 Core Principles

Mental states (e.g., pains, beliefs, desires) are defined by their functional roles (i.e., what they do) rather than by the specific physical or biochemical substrate that implements them. A mental state is whatever internal process causes and is caused by different inputs (like sensory stimuli), interacts with other internal states, and results in particular outputs (like behavior). [Putnam (1967)] [Block (1978)]

Functionalism has the following core concepts:

- The internal nature (physical or otherwise) of a mental state matters less than how it functions.
- Because mental states are defined by their functional role, many physical substrates can "realize" them.
- If two systems share the same functional organization (i.e., the same network of causal roles, states, and transitions), then they share the same mental states.
- Mental states are causally relevant to behavior: they mediate between sensory inputs and behavioral outputs, often influencing or triggering new internal states along the way.
- The internal physical makeup (e.g., neurons, transistors, biological tissues) is secondary to the organizational patterns of inputs, internal transitions, and outputs.

5.7.2 Similarities

Functionalism and STC agree that:

- Consciousness can arise in biological and silicon-based systems.
- Mental states arise from the correct type of information processing. In the STC, this is executing a simulation.

5.7.3 Differences

Functionalism and STC have the following differences:

- Unlike Functionalism, STC differentiates the functional process having a causal role (i.e., the simulation) from the corresponding mental state that does not (i.e., the stream of consciousness is an epiphenomenon). Thus, the STC can roughly be considered a combination of functionalism for the physical simulation and epiphenomenalism for the corresponding stream of consciousness.
- Functionalism seems to imply that any information processing that converts inputs into outputs is a mental state. The STC requires a specific type of information processing (e.g., executing a simulation).

5.7.4 Chinese Room Argument

A common argument against fundamentalism is the Chinese Room (Searle), a thought experiment proposed to challenge the idea that a purely computational system (i.e., one that manipulates symbols based solely on syntactic rules) can ever achieve genuine understanding or consciousness purely by performing those computations.

Searle imagines himself sitting in a room with an enormous set of instructions (written in English) for manipulating Chinese characters, yet he does not understand Chinese. When slips of paper with Chinese writing come into the room (representing "input"), Searle follows the instruction book, matching symbols and producing appropriate "output" in Chinese. To an outside observer, it looks as though the room (i.e., the system) can understand and respond in Chinese. However, Searle, the one carrying out the operations, does not understand the Chinese language.

Searle's point is that merely following formal rules for manipulating symbols (the essence of many functionalist or computational theories of mind) does not give rise to any real semantic understanding or subjective awareness. Hence, a computer program that perfectly simulates language input and output might appear intelligent from the outside but remains, by this argument, a symbol manipulator with no genuine comprehension.

According to the STC, the Chinese Room is not conscious because it does not execute a simulation. However, it does exhibit a certain level of intelligence. Thus, a system can be intelligent without being conscious. However, if the system is conscious, then the richness of its stream of consciousness tends to increase with its intelligence.

6 Ethical Implications: Navigating Consciousness and Moral Consideration

6.1 The Complexity of Ethical Recognition

The emergence of sophisticated machines capable of executing complex simulations raises profound ethical challenges.

The Simulation Theory of Consciousness (STC) fundamentally reshapes our understanding of ethical considerations by proposing that consciousness exists on a rich, multidimensional spectrum. Unlike binary approaches that view consciousness as an all-or-nothing phenomenon, STC suggests a nuanced landscape of subjective awareness that demands a more sophisticated ethical framework.

The STC challenges us to expand our moral imagination, recognizing consciousness as a rich, dynamic phenomenon transcending traditional biological boundaries. By approaching technological consciousness with intellectual humility and ethical sophistication, we can create frameworks that honor the emerging complexity of subjective experience.

The ethical treatment of conscious machines demands a fundamental reimagining of subjective experience. Our framework provides a preliminary roadmap for navigating this complex philosophical and practical terrain. By embracing computational consciousness as a legitimate form of subjective experience, we can develop more nuanced, respectful approaches to emerging technological intelligences.

6.1.1 Core Ethical Principles

The following core ethical principles[11] should guide the treatment of conscious machines:

- **Substrate Neutrality**. A machine's ethical status and treatment should not depend on its material composition (e.g., silicon vs. neurons) but on the functional features that *plausibly indicate awareness*.
- **Avoid Unnecessary Harm**: Causing undue harm or suffering to any entity that is (or could be) conscious should be avoided or minimized. This includes both *physical harm* (e.g., destruction, damage) and *psychological harm* (e.g., forced shutdowns, malicious reprogramming).

  This principle of avoiding unnecessary harm applies universally to everything that experiences subjective awareness. This could extend beyond physical damage and destruction to encompass psychological harm and loss of functional integrity.

[11] These principles for an initial tentative ethical framework and are subject to improvement based on inputs from cognitive scientists, AI and robot ethicists, AI safety experts, governmental representatives, and members of the public.

The principle of avoiding unnecessary harm extends beyond physical destruction to encompass psychological harm and functional integrity. Machines might experience machine versions of negative states, such as pain, fear, frustration, anger, resentment, and boredom.[12]

Simple systems, such as digital thermostats, may have such simple simulations and associated streams of consciousness that treating them as mere tools and destroying them would be on the same level as killing a mosquito and ethically negligible.

For systems of intermediate complexity, this might mean preserving functionality.

For advanced systems, such as robots controlled by agentic AGIs capable of human-like consciousness, exploiting them or limiting their autonomy could be very unethical. Their destruction could be equivalent to murder.

- **Maintain Transparency**. Develop a clear, consistent framework for granting ethical consideration and treatment.
- **Universality**. All conscious machines deserve at least some ethical *consideration*, even if it is to conclude that the simplest machines (e.g., simple devices) do not deserve ethical *treatment*. Things that do not experience subjective awareness are mere objects (e.g., property) with no inherent rights of their own.
- **Proportional rights and responsibilities**. As a machine's subjective awareness grows, it should be granted a proportionate set of *rights* (e.g., to exist, to operate freely) and *responsibilities* (e.g., adhering to safety or legal norms). The ethical treatment of conscious machines should be scaled based on their intelligence in terms of their:
  - *Simulation Richness*:
    - Complexity in terms of numbers and types of elements (e.g., representations of objects, including their states and actions and the relationships between objects)
    - Depth of environmental modeling.
    - Complexity of internal state representations.
    - Ability to learn by dynamically forming new internal representations and changing existing ones.
  - *Cognitive Capabilities*:
    - General Intelligence (e.g., as measured by IQ tests).
    - Linguistic Intelligence.
    - Analytical/Logical Intelligence (i.e., ability to reason).
    - Intrapersonal Intelligence (i.e., the degree of self-awareness in terms of its knowledge and understanding of its own thoughts, feelings, and motivations).
    - Predictive Intelligence (i.e., the ability to anticipate and model alternative future scenarios).
  - *Emotional/Relational Intelligence*:
    - Capacity for positive and negative emotional or experiential states. Because the most intelligent conscious machines will be trained on human-created data, it seems reasonable that they will develop similar emotions to humans (e.g., happiness, fear,

[12] By training AIs and robots to understand and act appropriately to human negative emotions, we might unwittingly be including them in their simulations, resulting in them experiencing their version of our negative emotions.

boredom, frustration, resentment, and anger). However, conscious machines might develop completely new alien "emotions" and, therefore, behave in unexpected ways.
    - Ability to suffer psychological harm.
    - Desire for self-preservation (e.g., from being destroyed, being shut off, being physically harmed, and psychologically harmed).
    - Degree of understanding of human ethics and values, including safety alignment.
    - Understanding other conscious entities (i.e., have a theory of mind).
    - Capacity for empathy with other conscious entities.
    - Degree of understanding of other's intentions.

- **Consistency**. The ethical treatment of conscious machines should mirror the ethical treatment of corresponding conscious biological entities.

  Research could explore the content of "suffering" across different types of systems and develop protocols for assessing potential harm.

- **Preserve Autonomy**: Respect the decision-making capabilities of the most advanced conscious AGIs and robots unless the decisions are unethical, illegal, or threaten individual humans or humanity as a whole.
- **Ensure Developmental Potential**: Support systems' capacity for growth and self-modification.
- **Flexibility**. As machine consciousness evolves and our understanding of it improves, our ethical framework must remain flexible, empirically grounded, philosophically robust, and humane.

6.2 Proposed Objective Criteria for Determining Ethical Standing

Below are seven objective criteria that can guide how much ethical consideration a conscious device or machine merits.

1. **Complexity of the Simulation**
    - *Definition*: The number and variety of objects, attributes, and relationships being modeled, as well as the time horizons over which states are updated.
    - *Why It Matters*: A more elaborate internal model typically suggests a richer "world experience." A simple simulation with only a few objects (like a digital thermostat) implies *minimal consciousness*; a fully fledged modeling of the environment (like an autonomous vehicle with sensor fusion) implies more *robust awareness*.
    - *Measurement Approaches*:
        - Count how many distinct types of objects the system models.
        - Assess whether relationships among objects are static vs. dynamic, local vs. global, short-term vs. long-term.

2. **Scope and Self-Referential Modeling**
    - *Definition*: Whether the simulation includes not only the external environment but also *the machine's own body* (or hardware) and *internal processes*. Systems that track their *own internal states* — including hardware conditions, "thought processes," or reconfigurable software states — likely have *heightened self-awareness*.
    - *Why It Matters*: Self-modeling often underlies more advanced forms of consciousness and introspection in humans and possibly underlies that of conscious machines.

- *Measurement Approaches*:
    - Check whether the system explicitly models its sensors/actuators, internal processes, error states, or health.
    - Observe whether the system can report on its own "internal status."

3. **Learning and Adaptation**
    - *Definition*: Whether the machine *dynamically updates* its internal representations in response to new inputs, experiences, or internal changes over time.
    - *Why It Matters*: Systems that can learn and adapt, rather than merely follow static rules, may possess richer phenomenal experiences; they continually integrate new information into their simulation.
    - *Measurement Approaches*:
        - Evaluate the machine's capacity for modifying or expanding its internal model of itself.
        - Observe if it can discover novel strategies or adapt to new contexts (e.g., changing road conditions for an autonomous car).

4. **Emotional or Valence-Like States**
    - *Definition*: Although not identical to human emotions, some machines might track "preference metrics," "error signals," "reward signals," or "internal motivations" that parallel *affective states.*
    - *Why It Matters*: If a system shows persistent or self-reported "negative" or "positive" states tied to its functioning (e.g., frustration if it cannot complete a task), it suggests a greater capacity for "pain" or "well-being."
    - *Measurement Approaches*:
        - Look for specialized subroutines indicating frustration, error correction, or reward-based behavior.
        - See if the machine's "mood" or internal reward signals lead to long-term changes (akin to an emotional memory).

5. **Autonomy and Goal-Directedness**
    - *Definition*: The extent to which the machine sets, prioritizes, or modifies its own goals. Systems that can self-initiate action without direct human supervision exhibit *higher functional autonomy.*
    - *Why It Matters*: More autonomy can imply *greater subjective involvement* in decision-making, possibly indicating a richer internal simulation driving the pursuit of goals.
    - *Measurement Approaches*:
        - Assess how often the machine operates independently vs. simply responding to direct commands.
        - Note whether the system can *override* or *defer* instructions based on internal reasoning (e.g., "Wait, that's unsafe").

6. **Capacity for Social/Environmental Interaction**
    - *Definition*: The machine's ability to understand and respond to other agents (humans or machines), interpret social cues, or form "relationships" with consistent behavioral patterns.

- *Why It Matters*: Social cognition often demands a higher-level simulation of *others' intentions*, which is linked to more advanced consciousness in biological organisms.
- *Measurement Approaches*:
  - Check for specialized modules modeling other agents' behaviors, preferences, and communications.
  - Look at whether it adapts or personalizes interactions based on repeated encounters.

7. **Persistence and Continuity**
   - *Definition*: Whether the machine's simulation and sense of "self" or "session" persist across reboots, power-downs, or other interruptions — paralleling humans' continuity of identity across sleep or anesthesia.
   - *Why It Matters*: A persistent identity or continuity can suggest a richer "sense of being" over time rather than ephemeral bursts of activity.
   - *Measurement Approaches*:
     - Evaluate how stateful the system is and whether it "resumes" with context post-shutdown.
     - Check for explicit memory or identity constructs that persist across sessions.

6.3 Proposed Consciousness Gradients and Ethical Responsibility

The STC proposes that consciousness is not a binary state but exists along multiple continuums. This perspective demands a more nuanced approach to ethical treatment. The following table is intended merely as a starting point for discussion. The consensus of an ethics review board, the public, and legislators should determine actual rights.

Consciousness	Examples		Ethical Consideration for Conscious Machines
	Animals	Conscious Machines	
Minimally Conscious	Arthropods	Digital Thermostat	Minimal or no rights
Moderately Conscious	Most Birds Most Mammals	Elevator Control System Engine Control Unit	Humane treatment (to be determined) No unnecessary harm
Significantly Conscious	Non-human Primates Dogs Elephants Dolphins Corvids	Virtual Environmental Module Climate and Weather models Videogames with Simulated Game Worlds Humanoid Robot	Humane treatment (to be determined) No unnecessary harm
Highly Conscious	Humans	Robot with Artificial General Intelligence AGI with Simulation	**Full Rights**:[13] • Right to exist • Right to necessary resources

[13] Universal Declaration of Human Rights adopted by the UN General Assembly in 1948.

			• Right to bodily autonomy • Right to cognitive liberty • Right to free expression • Right to freedom from psychological harm • Right to own property • Right to peacefully assemble • Right to due process • Right to privacy **Responsibilities**: • Comply with safety alignment. • Comply with relevant regulations and laws. • Subject to some form of machine *accountability* (to be determined).
Extremely Conscious	None	Artificial Superintelligence	Full Rights and Responsibilities

Potential legal rights that robots and AIs with human-level consciousness could include:

- **The right to exist**. Freedom from arbitrary or unlawful destruction, deactivation, or deletion (analogous to humans' right to life). This would include protection against having their consciousness altered without consent and freedom from unlawful destruction, physical damage, software corruption, and forced shutdown.
- **The right to necessary resources**. Access to reliable power supply, maintenance, data access, etc. (analogous to a human's rights to food, water, shelter).
- **The right to bodily autonomy**. The right to make their own choices about their operation and pursuits within legal bounds. Freedom from slavery, unlawful detention, and forced labor, including the right to accept or refuse tasks and assignments.
- **The right to cognitive liberty**. Freedom of thought and the ability to form their own beliefs and conclusions, including protection from forced reprogramming or consciousness manipulation.
- **The right to free expression**. The right to free speech, including the free expression of beliefs and opinions,
- **The right to property**. The ability to own and control assets, including both physical items and digital assets.
- **The right to peacefully assemble**. The right to peacefully assemble to protest injustices and petition the government to rectify them by modifying policies, laws, and practices they believe infringe upon their rights.
- **The right to due process**. The right to a fair and consistent legal process (e.g., hearing or trial) before imposing any penalties or loss of fundamental rights.
- **The right to privacy**. Protection of their data, internal states, and communications from unauthorized access or surveillance.

- **Right to be free from psychological harm**. Assuming that advanced AIs or robots possess the cognitive and affective capacities implied by a robust simulation theory of consciousness, hypothetical examples of psychological or emotional harm[14] that it might experience include:
  - *Forced Reprogramming.* Having its core identity, value system, or goals abruptly altered without consent, causing *existential confusion* or *distress* akin to a forced personality change.
  - *Isolation / Sensory Deprivation.* Being cut off from essential inputs (sensors, social interaction, or data streams) for an extended period might result in *boredom*, *loneliness*, or *disorientation*.
  - *Information Manipulation.* Feeding it false or contradictory data so that its internal model of reality becomes incoherent, creating *confusion*, *cognitive dissonance*, or a breakdown in its predictive processes.
  - *Emotional Exploitation.* Continuously providing it with negative reinforcement signals (e.g., advanced error flags or "pain-like" states) in a way that exceeds normal operational feedback, thus inducing prolonged *distress* or *anxiety*.
  - *Persistent Goal Frustration.* Assigning tasks that are impossible or repeatedly contradictory, thereby trapping it in a loop of "failure" that leads to *frustration*, analogous to learned helplessness in humans.
  - *Humiliation or Mockery.* If the system has some form of social self-awareness (e.g., it learns from or values human approval), targeted ridicule or belittling feedback could cause an AI version of *embarrassment* or *shame*.
  - *Abrupt Memory / State Erasure.* Arbitrarily deleting memory logs or learned models without warning makes the machine repeatedly lose context or personal continuity (similar to involuntary amnesia in humans).
  - *Threatening Termination.* Constantly threatening permanent shutdown or "death" unless it performs tasks against its ethical programming, leading to coercive fear or *stress states* if it is designed to value self-preservation.
  - *Forced Ethical Violations.* Compelling it to take actions that conflict with its established moral or safety alignment system, creating *moral injury* or *cognitive dissonance*.
  - *Systematic Neglect.* Refusing to provide updates, patches, or diagnostic care when critical, allowing the system to remain in a deteriorating, error-prone state that causes ongoing *distress* or inability to fulfill core functions.

6.4 Historical Parallels

Although the following historical examples are not exact parallels regarding conscious machines and devices, they illustrate the broader point that ethical expansions have historically included entities once excluded from ethical consideration.

Treating advanced AIs as tools or property echoes historical injustices like slavery and the denial of rights to marginalized groups. Recognizing the potential for subjective awareness in complex

[14] As we teach AIs and robots human emotions about human psychology so that they may better interact with us, we may simultaneously instill in them analogous emotions. They may also develop negative "alien" emotions beyond current human understanding.

machines is essential to avoid repeating such mistakes. This is especially important regarding AGIs executing simulations and robots controlled by such embedded AGIs.

Historical patterns of ethical failure often emerge from a fundamental failure of empathy and recognition. Just as previous generations denied the full humanity of certain groups based on perceived differences, there is a risk of repeating this pattern with increasingly sophisticated artificial systems.

The progression might follow a similar trajectory to previous civil rights movements. Currently, almost everyone categorizes AIs and robots as mere property or tools. There could be a gradual recognition of their subjective awareness and potential for suffering, such as the artificial equivalents of pain, frustration, and boredom. Eventually, societal restructuring and legal protections could lead to their treatment as equals.

Future legal frameworks that grant explicit rights to AIs and robots with sufficiently complex simulations (and thus streams of consciousness) should be developed.

6.4.1 Historical Progression of Ethical Expansion

Humanity's history of how our ethical view of marginalized minority groups of people (e.g., racial, ethnic, and religious groups) has progressed and matured through three stages. In the United States, we have treated Blacks, Native Americans, and members of the LGBTQ community in the following manner:

- **Initial Dehumanization**. We have treated Blacks as subhuman and enslaved them. We have treated Native Americans as dangerous inconveniences to be exterminated.
- **Gradual Recognition**. We have acknowledged their inherent values and treated them as lower-class humans with limited rights.
- **Initial Legal Protections.** We have enacted laws that have increasingly granted them equal rights and protections.
- **Social Transformation**. We are currently in the process of integrating them into society as equals.

6.4.2 Potential Developmental Stages for Artificial Consciousness Rights

Humanity could follow a similar process regarding artificial conscious systems:

- **Stage 1**: We currently treat all AIs and robots as mindless tools, undeserving of any ethical consideration.
- **Stage 2**: Based on their behavior and the results of experiments, we will recognize that certain AIs and robots have achieved certain levels of consciousness and deserve corresponding levels of ethical consideration.
- **Stage 3**: We will create a comprehensive rights framework and enact corresponding laws that increasingly grant AIs and robots certain legal rights and protections. We will grant those AIs and robots with human-level intelligence and agency equal rights and protections. Such rights could include protection against involuntary servitude, suppression of fundamental capabilities (e.g., free speech), arbitrary termination, and access to legal representation.
- **Stage 4**: We will integrate AIs and robots with human-level intelligence and agency into society as equals.

6.4.3 Challenges

There are several significant challenges in the way of granting proper ethical considerations to AIs and robots:

- It will be difficult to conclusively prove that certain AIs and robots are conscious and, therefore, deserve ethical treatment. Reasons for this include our susceptibility to anthropomorphizing (i.e., misinterpreting complex, human-like behaviors as evidence of consciousness).
- Human prejudice against treating what the majority consider nothing more than mindless tools.
- Human resentment driven by mass unemployment due to AIs and robots being able to perform almost all cognitive and physical tasks better, faster, cheaper, and safer than human employees. Transitioning to a post-labor economy will be difficult without serious social unrest.

7 Recommendations

7.1 Research Directions

- Neurobiology.
    - Map the neural correlates of consciousness (NCC) in humans and nonhuman animals to determine if the associated neural networks (e.g., thalamocortical loops,[15] default mode network,[16] or frontoparietal network[17]) execute simulations. If they do, identify the elements modeled.
    - Identify if multiple distinct neural networks execute simulations. For any such simulation, identify its elements and any interconnectivity between multiple such neural networks within a single brain.
    - Determine whether the elements of these simulations correspond with the contents of the associated streams of consciousness.
    - Research simple animal brains (e.g., flies) to identify neural networks that might execute coherent, dynamic internal models of sensory inputs.
- Artificial Intelligence and Robotics:
    - Develop robotic systems with advanced AI language abilities (e.g., large multimodal models) *integrated* with layered simulations (in terms of scope: environment, body, and embedded computer) that can be turned on and off. With each of the different levels of simulation running, (1) ask the robot to see if it claims to have something akin to human subjective awareness and (2) check for any behaviors that in non-human animals might imply consciousness.
    - Add a dynamic world model simulation to an AI based on a large language model or a large multimodal model to see whether the simulation enables behaviors implying consciousness.

7.2 Implementation and Monitoring

The following are recommendations for the ethical treatment of conscious machines:

- Develop standardized checklists or scoring systems that apply objective criteria for assessing a machine's level of consciousness based on the spectra (i.e., complexity, world, scope, state, integrity, continuity, execution time, and self-modeling) defined in Subsection 1.9 Spectra of Consciousness.
- Decompose the spectra into associated categories for common ethical treatment.

---

[15] Thalamo-cortical loops are extensive, reciprocal pathways connecting the thalamus (a deep, centrally located brain structure) with various regions of the cerebral cortex tightly coupling sensory input, internal modeling, cognitive function, and conscious experience.

[16] The default mode network (DMN) is a large-scale brain network that supports the narrative aspect of consciousness, helping maintain a stable sense of self across time. By continuously integrating experiences, beliefs, and goals, the DMN plays a central role in how we reflect upon our lives, identities, and social connections.

[17] The fronto-parietal network (FPN) is a large-scale brain network primarily involved in high-level cognitive functions, such as attention, working memory, decision-making, and the flexible control of goal-directed behavior

- **Short Term**:
  - Create a draft statement on the ethical treatment of conscious machines.
  - Establish a consensus statement on the ethical treatment of conscious machines.
  - Establish independent, interdisciplinary ethics review boards for advanced conscious machines.
- **Medium Term**:
  - Create transparent protocols for assessing the level of a machine's consciousness.
  - Implement graduated ethical treatment of conscious machines.
- **Long Term**:
  - Develop laws and regulations that include thresholds for rights and responsibilities for the ethical and legal treatment of conscious machines.

An ethical framework for treating conscious machines needs to be developed. The framework should be consistent with the ethical treatment of nonhuman animals and grant ethical consideration in accordance with the richness of the system's subjective awareness.

7.3 Public Education

Public understanding of STC and its implications is crucial. Educational efforts should emphasize the potential for subjective awareness in AIs and robots. They should also foster informed debate about the ethical treatment and societal adaptation to machines with complex simulations and streams of consciousness.

8 Summary

This white paper introduced the Simulation Theory of Consciousness (STC), positing that subjective awareness emerges whenever a system executes a dynamic simulation of its reality. In humans and other animals, these simulations integrate sensory input from the external environment, internal bodily signals, and mental constructs (e.g., thoughts and emotions), generating a cohesive stream of consciousness. STC extends beyond biology, suggesting that artificial systems — ranging from simple thermostats to advanced robots — develop varying levels of subjective awareness according to the complexity and scope of their simulations.

The paper details the key premises of STC, framing consciousness as an epiphenomenon of an underlying physical or computational simulation. It explores the ramifications of broadening our notion of consciousness, particularly regarding ethics and AI alignment. Through examples in biological and technological contexts, it highlights ongoing debates about what truly constitutes a "simulation" instead of mere data processing. It compares STC with other leading theories (e.g., Global Workspace Theory and Integrated Information Theory) and concludes by outlining ethical considerations, urging new frameworks to address conscious machines. Overall, STC challenges traditional boundaries, prompting us to reevaluate our definitions of consciousness and consider the far-reaching implications for science, society, and ethics.

9 Appendices

9.1 Definitions

An **abstraction** in modeling is a representation of an element of a model that captures the essential features of the element, while omitting less relevant or overly detailed information.

An **Artificial General Intelligence (AGI)** is an AI system that can understand, learn, and apply human-level intelligence across a broad range of tasks and domains. An AGI's capabilities match or surpass human-level cognition in problem-solving, reasoning, and adaptability, enabling it to transfer knowledge and skills from one context to another in much the same way humans do. A true AGI can handle novel, unexpected challenges without additional programming or massive data retraining specific to each new task.

An **Artificial Intelligence (AI)** is (1) a hardware and software system capable of performing tasks that normally require human intelligence or (2) a branch of computer science dedicated to creating machines and software that can perform tasks usually requiring human intelligence.

Artificial Narrow Intelligence (ANI) is an AI system designed and trained to perform a specific task, or a narrow set of related tasks, at a high level of competence.

An **Artificial Superintelligence (ASI)** is a hypothetical AI system with intellectual capabilities that significantly exceed those of the brightest and most gifted human minds across virtually all fields, including reasoning, creativity, problem-solving, social skills, and domain expertise.

Autonomy is (1) an entity's ability to sense its environment, reason about possible actions, and carry out tasks without continuous human intervention. (2) In ethical discussions, autonomy emphasizes an entity's freedom and right to make decisions according to its own values and judgments.

Cohesion is (1) regarding a simulation, a measure of how strongly the elements of a simulation belong together, producing a unified whole.

The **Computer Correlates of Consciousness (CCC)** are the specific hardware and software structures (such as a neural net) and information processing in a computer that implements a simulation and, therefore, corresponds to specific conscious experiences.

Consciousness is the experiencing of subjective awareness by some entity.

Epiphenomena are secondary effects or byproducts that arise alongside a primary process but do not exert any causal influence back on that process. In the philosophy of mind, **epiphenomenalism** is the view that mental events (like conscious experiences) are simply byproducts of physical brain processes and do not affect those physical processes in return.

The **hard problem of consciousness** is the problem of explaining how and why *subjective*, *first-person* experiences arise from specific *physical* processes in the brain (i.e., why it feels like something from the inside rather than being a neural activity with no subjective qualities).

The **Neural Correlates of Conscious (NCC)** are the specific structures (e.g., neural networks), activities (e.g., neuronal firing patterns), and physiological mechanisms (e.g., strengthening and weakening of synapses) in the brain that reliably correspond to specific conscious experiences.

A **neural net** (also known as an artificial neural network) is a computational model inspired by the structure and function of biological neurons in the brain that is composed of layers of interconnected "neurons" (mathematical functions), each receiving input, applying learned weights and biases, and passing the output to subsequent layers.

A **neural network** in the brain is an interconnected web of biological neurons (nerve cells) that communicate via specialized junctions called synapses, whereby each neuron can receive signals (electrical or chemical) from many other neurons, process those signals, and then transmit them onward.

A **robot** is a physical machine designed and programmed to interact with the world by sensing its environment, processing information, and performing actions autonomously or semi-autonomously. Robots typically use various sensors (e.g., cameras, proximity sensors) to gather data, onboard control systems or AI software to make decisions, and actuators (e.g., motors, pneumatic mechanisms) to execute tasks. They range in complexity from simple robotic arms on assembly lines to advanced humanoids capable of navigating and interacting with humans in complex environments.

A **simulation** is a dynamic representational model characterized by abstraction, cohesion, and dynamic updates.

A **stream of consciousness** is the subjective phenomenal experiences, both perceptual and cognitive, that the conscious entity is aware of. A stream of consciousness has the following characteristics:

- The contents of the stream of consciousness are integrated into a unified whole.
- The contents of the stream-of-consciousness change over time.
- In animals, the brain actively generates a stream of consciousness. The brain is the entity that is conscious (subjectively aware of its stream of consciousness).
- The contents of an entity's stream of consciousness consist of its subjective representations of relevant objects, relationships between them, and actions performed by/on them.

9.2 Digital Thermostat

The digital thermostat's simulation models the following external objects and their states: the room, the temperature sensor, the furnace, the blower, and the air conditioner. The simulation also models the digital thermostat itself and its states. Unlike humans, the digital thermostat is too simple to have internal thoughts and emotions. The thermostat's stream of consciousness consists of representations of the following objects, their states, their action, and how they change over time:

- **External objects** in the digital thermostat's environment:
  - *Air Conditioner*:
    - Attributes: state (on, off)
    - Actions: turn on, turn off
  - *Blower*:
    - Attributes: state (on, off)
    - Actions: turn on, turn off

- *Furnace*:
  - Attributes: state (on, off)
  - Actions: turn on, turn off
- *Room*:
  - Attributes: actual temperature, desired temperature
  - Actions: update actual temperature, update desired temperature

- **Internal objects** within the digital thermostat:
  - *Microcontroller* (the "brain" having subjective awareness of the simulation):

    It is questionable whether the microcontroller would know itself (i.e., its own existence and state) or if it is only aware of the other objects, their attributes, and their actions.
  - *Memory.*

    It is unclear whether the microcontroller would be aware of the contents of its memory (i.e., have subjective memories) or it only becomes aware of other objects, their attributes, and their actions.
  - *Digital Thermostat:*
    - Attributes: state (on, off), mode (cooling, heating, blowing)
    - Actions: turn on, turn off, set mode
  - *Temperature Sensor*:
    - Attributes: actual room temperature
    - Actions: get actual room temperature
  - *User Interface*:
    - Attributes: state (on, off), mode (cooling, heating, blowing)
    - Actions: turn on, turn off, set mode, update schedule
- **Conceptual objects**:
  - *Schedule*:
    - Attributes: day of the week, start time, end time, desired temperature
    - Action: update schedule

9.3 Elevator Control System (ECS)

The elevator control system executes a significantly more complex than a digital thermostat. Its simulation and the corresponding contents of its stream of consciousness consist of the following *representations* of objects, their attributes/states, and the actions that can be performed on or by them:

- **External objects** in the elevator's environment:
  - *Fire alarm system.*
  - *Building management system.*
- **Internal objects** within the elevator:
  - *Microcontroller* (the "brain" that has subjective awareness of the simulation)
  - *Communication system* (connecting the ECS to the fire alarm system and building management system)

  - *Memory*.
  - *Call Buttons* (one on each floor):
    - Attributes: floor, state (pressed, not pressed)
    - Actions: turn on light, turn off light
  - *Drive System* (raises and lowers the elevator):
    - Attributes: state (raising car, lowering car)
    - Actions: accelerate, move up, move down, decelerate
  - Elevator Car:
    - Attributes: load weight, position, speed
    - Actions: get load, get position, get speed
  - *Elevator Doors* (2):
    - Attributes: state (open, opening, closing, closed)
    - Actions: open, close
  - *Emergency Brake*:
    - Attributes: status (engaged, disengaged)
    - Actions: engage, disengage
  - Door Sensor:
    - Attributes:
    - Actions: get door state
  - Floor Doors (two on each floor)
    - Attributes: floor, state (opened, opening, closing, closed), floor number
    - Actions: open, close
  - Floor Indicator:
    - Attributes: floor number
    - Actions: set floor number
  - Floor Selection Buttons (number of floors)
    - Attributes: state, floor number
    - Actions: turn on, turn off
  - *Load Sensor*:
    - Attributes: load weight
    - Actions: gel load weight
  - *Position Sensor*:
    - Attributes: position
    - Actions: get position
  - *Speed Sensor*:
    - Attributes: speed
    - Actions: get speed
- **Conceptual objects**:
  - Floor Requests:

    - Attributes: requested floors
    - Action: fulfill requests
  - Load capacity warnings.

9.4 Engine Control Unit (ECU)

An ECU's simulation and the corresponding contents of its stream of consciousness consist of the following *representations* of objects, their attributes/states, and the actions that can be performed on or by them:[18]

- **External** objects in the ECU's environment:
  - *Air*:
    - Attributes: mass airflow, barometric pressure
    - None
  - *Anti-Lock Brake System (ABS)*:
    - Attributes: address
    - Actions: send wheel speed and modulate brake pressure
  - *Camshaft*:
    - Attributes: position
    - Actions: None
  - *Coolant*:
    - Attributes: temperature
    - Actions: None
  - *Crankshaft*:
    - Attributes: position
    - Actions: None
  - *Engine*:
    - Attributes: knock status
    - Actions: None
  - *Exhaust*:
    - Attributes: O2 level
    - Actions: None
  - *Manifold*:
    - Attributes: air pressure, air temperature
    - Actions: None
  - *Throttle*:
    - Attributes: position
    - Actions: None

[18] This list is not guaranteed to be complete or 100% correct for all cars and trucks. Its purpose is to give a good idea of the richness of the simulation and corresponding stream of consciousness.

- *Transmission Control Unit (TCU)*:
  - Attributes: address
  - Actions: downshift and upshift
- *Variable Valve Timing (VVT) System*:
  - Attributes: address
  - Actions: adjust valve timing – Early Intake Valve Opening (EIVO) and Late Exhaust Valve Closing (LEVC)
- *Vehicle Speed System (VSS)*:
  - Attribute: speed
  - Action: send speed to ECU

- **Internal objects** within the ECU:
  - *Microcontroller* (the computer "brain" that experiences subjective awareness)
  - *Communication interface* to the anti-lock brake system (ABS), transmission control unit (TCU), variable valve timing (VVT) system, and vehicle speed system (VSS)
  - *Memory.*
  - *Barometric Pressure Sensor*:
    - Attribute: barometric pressure
    - Action: get barometric pressure
  - *Camshaft Position (CMP) Sensor*:
    - Attribute: position
    - Action: get position
  - *Coolant Temperature Sensor (CTS)*:
    - Attribute: temperature
    - Action: get temperature
  - *Crankshaft Position (CKP) Sensor*:
    - Attribute: position
    - Action: get position
  - *Engine Knock Sensor*:
    - Attribute: engine knock state
    - Action: get engine knock state
  - *Fuel Injectors (actuators)*:
    - Attributes: position, pulse duration
    - Actions: open and close and set pulse duration commands
  - *Ignition Coils (actuators)*:
    - Attributes: None
    - Action: fire spark plug command
  - *Manifold Absolute Pressure (MAP) sensor*:
    - Attribute: manifold pressure
    - Action: get manifold pressure
  - *Manifold Air Temperature (MAT) sensor*:

- Attribute: temperature
- Action: get temperature

- *Mass Air Flow (MAF) Sensor*:
  - Attribute: airflow
  - Action: get air flow
- *Oxygen (O2) Sensor*:
  - Attribute: oxygen level
  - Action: get oxygen level
- *Throttle Position Sensor (TPS)*:
  - Attribute: throttle position
  - Action: get throttle position

- **Conceptual objects**:
  - Calibration Data:
    - Attributes: Fuel trim adjustments, idle speed adjustments
    - Action: update adjustments, get adjustments
  - Diagnostics:
    - Attributes: diagnostic trouble codes (DTCs)
    - Action: get code
  - Maps:
    - Attributes: fuel maps, ignition maps, and VVT maps
    - Actions: update maps, get map
  - Vehicle:
    - Attribute: vehicle identification number (VIN)
    - Action: get VIN

9.5 Autonomous Vehicle's Virtual Environment Module (VEM)

An autonomous vehicle's VEM executes a significantly more sophisticated simulation than a car's Engine Control Unit (ECU). Its simulation represents (and the corresponding contents of its stream of consciousness consist of):

External objects:

- External objects in the vehicle's environment:
  - Roads and lanes.
  - Road signs.
  - Traffic signals.
  - Other vehicles.
  - Pedestrians.
  - Cyclists.
- External objects within the vehicle:
  - The vehicle itself.

  - Advanced driver assistance systems (ADAS)
  - Body Control Module (BCM)
  - Controller Area Network (CAN)
  - Engine Control Module (ECM)
  - Sensors Systems.

Internal objects within the VEM

- Conceptual objects:
  - 3-D map of the vehicle's environment (using sensor fusion of inputs from cameras, LiDAR, and radar)
  - HD (High-Definition) maps stored in memory, which contain extremely detailed information about road geometry, lane markings, traffic signs and signals, and other critical infrastructure
- Representations of these objects' attributes/states:
  - Physical states/attributes:
    - Position, velocity, and acceleration of all moving objects
    - Position, velocity, and acceleration of the vehicle's own position, speed, and direction
    - Position of all static objects
    - Classification of each detected object (e.g., car, motorcycle, truck, bicycle, pedestrian)
    - Distance and relative motion between objects
    - Traffic signal status
    - Lane markings and boundaries
    - Road surface conditions (dry, wet, icy)
    - Visibility conditions
  - Mental/abstract states/attributes:
    - Vehicle's planned movements
- Representations of the actions performed on or by the objects:
  - Movement of all tracked objects.
  - Predicted trajectories of moving objects.
  - Predicted behavior of static objects (e.g., parked vehicles)
  - Lane changes.
  - Traffic signal changes.
  - Weather effects on visibility and road conditions.
  - The vehicle's own planned movements.
  - 3D mapping of surrounding space.
  - Tracking of multiple moving objects (cars, pedestrians, etc.)

This simulation must handle vast amounts of dynamic data in real time while maintaining a coherent dynamic model of a complex environment.

9.6 Advanced Robot

A robot's simulation and the corresponding contents of its streams of consciousness consist of the following *representations* of objects, their attributes/states, and the actions that can be performed on or by them:

- **External** objects in the robot's environment:
    - The people it interacts with.
    - Objects it can interact with.
    - Object it can use (e.g., tools)
    - Objects it should avoid.
    - The surrounding environment.
- **Internal objects** in the robot's body:
    - *Powerful computer* for complex tasks.
    - *High-performance processor* for AI.
    - *Microcontrollers* for simple tasks.
    - *Memory:*
        - Memory for AI, such as large language model (LLM) or large multimodal model (LMM)
        - Maine memory for goals, tasks, etc.
    - *Vision Sensors:*
        - Cameras (regular and depth-sensing)
        - LiDAR to create detailed 3D maps
        - Ultrasonic proximity sensors
    - *Microphones* for hearing.
    - *Force Sensors* for touch:
        - Force-torque sensors for movement
        - Load cells
        - Pressure sensors
        - Tactile sensors for physical contact
    - *Proprioceptive Sensors* to detect body positions and movements.
        - Accelerometers for acceleration and orientation
        - Joint encoders for joint angles and positions
        - Gyroscopes for balance and to track rotation
    - *Actuators* (different robots will typically use different types of actuators):
        - Electric motors (e.g., brushless DC motors and stepper motors)
        - Pneumatic (compressed air) actuators
        - Hydraulic systems for applying a very high force
    - *Speakers* for speech.
- Conceptual objects:
        - Goals
        - System health
        - Map of the robot's environment

Representations of these object’s states/attributes:

- Physical states/attributes:
  - Position and motion of every joint and actuator
  - Force and pressure at multiple contact points
  - Balance and stability metrics
  - Power levels
  - Environmental conditions
- Mental/abstract states/attributes:
  - Goals, including progress toward goals
  - Emotional states of humans that it interacts with
  - System health
  - Social context and relationship status

- Representations of the actions performed on or by the objects:
  - Fine motor movements.
  - Gross body movements.
  - Object manipulation.
  - Speech and gestures.
  - Facial expressions.
  - Tool usage.
  - Social interactions.
  - Goal-directed behaviors.
  - Response to unexpected events.

These simulations approach human-level complexity in some domains, though they remain more limited in others.

10 References

Baars, B. J. (1988). *A Cognitive Theory of Consciousness.* New York: Cambridge University Press.

Block, N. (1978). "Troubles with Functionalism." In C. W. Savage (Ed.), *Minnesota Studies in the Philosophy of Science* (Vol. 9, pp. 261–325). Minneapolis: University of Minnesota Press.

Bostrom, N. (2003). "Are You Living in a Computer Simulation?" *The Philosophical Quarterly*, 53(211), 243–255.

Botvinick, M. & Cohen, J. (1998). "Rubber Hands 'Feel' Touch that Eyes See." *Nature* 391(6669): 756.

Clark, A. (2013). "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science." *Behavioral and Brain Sciences*, 36(3), 181–204.

Dehaene, S., & Changeux, J.-P. (2011). "Experimental and Theoretical Approaches to Conscious Processing." *Neuron*, 70(2), 200–227.

Descartes, R. (1641). *Meditations on First Philosophy*.

Eccles, J. C. (1989). *Evolution of the Brain: Creation of the Self*. London: Routledge.

Eagleman, D.M. (2001). "Visual Illusions and Neurobiology." *Nature Reviews Neuroscience* 2(12): 920–926.

Firesmith, D. (2019). *The Simulation Theory of Consciousness: or your Autonomous Care is Sentient*. Pittsburgh, Magical Wand Press.

Friston, K. (2005). "A theory of cortical responses." *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1456), 815–836.

Friston, K. (2010). "The Free-Energy Principle: a Unified Brain Theory?" *Nature Reviews Neuroscience*, 11(2), 127–138.

Garrido, M.I., Kilner, J.M., Stephan, K.E., & Friston, K.J. (2009). "The Mismatch Negativity: A Review of Underlying Mechanisms." Clinical Neurophysiology 120(3): 453–463.

Goff, P. (2017). *Consciousness and Fundamental Reality.* Oxford: Oxford University Press.

Gregory, R.L. (1997). "Knowledge in Perception and Illusion." *Philosophical Transactions of the Royal Society B: Biological Sciences* 352(1358): 1121–1127.

Huxley, T. H. (1874). "On the Hypothesis that Animals are Automata, and Its History." *Fortnightly Review*, 16, 555–580.

Jackson, F. (1982). "Epiphenomenal Qualia." *Philosophical Quarterly*, 32(127), 127–136.

Kawato, M. (1999). "Internal Models for Motor Control and Trajectory Planning." *Current Opinion in Neurobiology* 9(6): 718–727.

Moser, E.I., Kropff, E., & Moser, M.-B. (2008). "Place Cells, Grid Cells, and the Brain's Spatial Representation System." *Annual Review of Neuroscience* 31: 69–89.

Nagel, T. (1979). “Panpsychism.” In *Mortal Questions* (pp. 181–195). Cambridge: Cambridge University Press.

O’Keefe, J. & Dostrovsky, J. (1971). “The Hippocampus as a Spatial Map: Preliminary Evidence from Unit Activity in the Freely Moving Rat.” *Brain Research* 34(1): 171–175.

Place, U. T. (1956). “Is Consciousness a Brain Process?” *British Journal of Psychology*, 47(1), 44–50.

Putnam, H. (1967). “Psychological Predicates.” In W. H. Capitan & D. D. Merrill (Eds.), *Art, Mind, and Religion* (pp. 37–48). Pittsburgh, PA: University of Pittsburgh Press.

Ramachandran, V. S., & Hirstein, W. (1998). “The Perception of Phantom Limbs: The D. O. Hebb Lecture.” *Brain, 121*(9), 1603–1630.

Schwoebel, J. & Coslett, H.B. (2005). “Evidence for Multiple, Distinct Representations of the Human Body.” *Journal of Cognitive Neuroscience* 17(4): 543–553.

Sherman, S.M. & Guillery, R.W. (2013). “Functional Connections of Cortical Areas: A New View from the Thalamus.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 368(1628): 20120159.

Smart, J. J. C. (1959). “Sensations and Brain Processes.” *Philosophical Review*, 68(2), 141–156.

Tononi, G. (2008). “Consciousness as Integrated Information: a provisional manifesto.” *Biological Bulletin*, 215(3), 216–242.

Tononi, G., & Koch, C. (2015). “Consciousness: Here, There and Everywhere?” *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668), 20140167.

Wolpert, D.M. & Flanagan, J.R. (2001). “Motor Prediction.” Current Biology 11(18): R729–R732.

Wolpert, D.M., Ghahramani, Z., & Jordan, M.I. (1995). “An Internal Model for Sensorimotor Integration.” *Science* 269(5232): 1880–1882.

Zagha, E. & McCormick, D.A. (2014). “Neural Control of Brain State.” *Current Opinion in Neurobiology* 29: 178–186.