Getting Beyond the Mind-Body Problem:

Scientific Phenomenism

Dale O. Stahl

ISBN: 9798860171664

DEDICATION

I dedicate this book to my parents who encouraged me to always do my best and be kind, to my wife who has stood by me despite the many times I've been lost in thought, and to my many mentors.

I am especially indebted to Mike Pore and the philosophical discussion group of the Austin UU Fellowship who have inspired me to write this book.

CONTENTS

I. Introduction

The mind-body problem has stymied philosophy and science for at least 400 years. It is commonly attributed to Descartes' (1641) assertion that mental things (such as conscious experiences and ideas) are different in kind from physical things (such as rocks, plants and animals). How mental things and physical things interact (if at all) remains a mystery. Science focuses exclusively on publicly observable/measurable things (a.k.a. physical things), and since conscious experiences are purely private, they fall outside the purview of science. While neuroscience deals with observable/measurable brain processes, the assertion of an identity between brain processes and conscious experiences remains controversial.

To get beyond the mind-body problem, we first need to understand it and why it is still an unresolved problem. Therefore, Chapter II presents a critique of four schools of thought on the

problem: dualism, physicalism, mentalism and phenomenism[1]. The take-away is that dualism, physicalism and mentalism have serious problems while phenomenism is a potential path forward. Chapter III delves more deeply into phenomenism. Critical to this task is an understanding of the concept of a *model* of reality, the essential role of those models in interpreting conscious experiences and the essential role of conscious experiences in modifying those models. As a by-product, the hard problem of consciousness is dissolved. The conclusion is that *Scientific Phenomenism*, as developed here, gets us beyond the mind-body problem.

Nonetheless, there remains a glaring shortcoming: neither science nor phenomenism provides an adequate answer to what a *self* is. Chapter IV tackles this shortcoming by proposing that my self is an object in a model of phenomenal reality, and further that I have a hierarchy of models of my self. The concept of a model is crucial to both resolving the mind-body problem and understanding what a self is.

Chapter V argues that Scientific Phenomenism provides a consistent interpretation of Quantum Mechanics as a model of phenomenal reality, thereby resolving the controversial *measurement problem*. Finally, Chapter VI explores implications of scientific phenomenism beyond the mind-body problem.

[1] The original term was "phenomenalism", but I prefer the shorter and more recent term "phenomenism"; e.g. see Brrne (2004).

II. A Critique of the Schools of Thought on the Mind-Body Problem.

The prominent schools of thought on the mind-body problem are dualism, physicalism, mentalism and phenomenism. Each of these will be described and critiqued below.

A. Dualism.

Dualism appeals to our common-sense view that conscious perception of a physical object is different from the physical object itself. For instance, we cannot touch and push our thoughts around (except metaphorically) as we can touch and push physical objects. This common-sense view is part of our inheritance from ancient times when humans conceived of gods who intervened via physical force but were not themselves governed by the same natural laws - thus different in kind from the objects one could touch and push.

After an exhaustive process of doubting his beliefs about the physical world, Descartes (1641) reached the seminal conclusion that “I think, therefore I am”. In other words, since he can deny absolute knowledge about the physical world but he is absolutely certain that he is having conscious deliberations, the physical and the mental are different-in-kind. Note that the famous statement “I think, therefore I am” is actually a trivial syllogism: the conclusion is subsumed in the premise. Almost any statement of the form “I (verb), therefore I am” would serve the same purpose. It is not the thinking that is essential; any conscious experiencing (such as seeing, hearing, etc.) would suffice. The contrapositive is “if I do not exist, then there is no verb and instance for which ‘I (verb)’ is true.”

Following Descartes, dualists assert that all things can be classified as consisting of one of two kinds of substance: mental and physical. This expression of dualism is currently called “substance dualism”. An alternative expression is “property dualism” which asserts that there is just one kind of substance but two kinds of properties. The latter is a semantic distinction without a difference; it still divides all things into the same two classes albeit with different names.

The main challenge to dualism is the supposed interaction between the physical and the mental. If there is no interaction, then the physical and mental are completely separate worlds (spaces), and our thoughts about physical things have no relation

whatsoever to the physical world; i.e. there is no mind-body problem. On the other hand, some dualists assert that coincidentally the mental and physical are in a one-to-one "parallel" relationship, but then the mental is a redundant representation of the physical, so there is no mind-body problem.

Dualists (like physicalists) assume there is a causal link going from the physical to the mental. For the 400 years since Descartes, there have been tremendous advances in science in general, but all our advances in neuroscience have stopped well short of demonstrating the necessity of the conscious experience that accompanies neuro activity. Quite to the contrary, neuroscience has succeeded in verifying the causal link between the neuro processing of our senses and the neuro activity of the behavior that follows, all without needing a role for mental things such as consciousness. It appears that a comprehensive theory of human behavior does not need the concept of mental things.

In addition, some dualists assume there is also a causal link going from mental processes to physical processes. This view is called *two-way causation*. It is a widely held (some would say self-evident) belief that our thoughts and feelings influence how we behave. But how can something non-physical influence something physical? Such an interaction violates the laws of physics in which every interaction involves the exchange of mass and/or energy. Therefore, for two-way causation to be true, current physics must be false. While physics 2500 years from now

will likely be quite different from our current understanding, just as our current physics is quite different from Aristotelian physics, nevertheless, given that the domain of physics will forever remain the objectively observable and measurable properties of things, the strictly private nature of conscious experiences will preclude a testable explanation for conscious experiences. In conclusion, dualism produces the mind-body problem rather than resolving it.

B. Physicalism (Materialism)

Materialism, which holds that everything is material, can be dated back at least to Democritus (400 BCE). It rose in stature following the successes of Newtonian physics. In the 1930s, Neurath (1931) and Carnap (1932) introduced the term "physicalism" to refer to an updated version that replaces "material" with "mass and/or energy"; i.e. to be *physical* means to have mass and/or energy as defined by contemporary physics. Thus, *physicalism* holds that everything is physical, implying that all so-called mental things are in fact physical.

What does "everything is physical" mean? To answer this, first we must understand what it means to be physical. Specifically, we need (i) a set of sufficient verifiable properties for something to be *physical*, and (ii) a method for deciding whether or not any particular thing has those properties. For example, in ancient times to be physical one had to be able to touch and feel it. In modern times, to be physical requires that the thing has mass and/or energy according to the contemporary scientific definition of mass and

energy, and must obey the laws of contemporary physics. The methods for deciding whether a thing is physical are given by prescribed procedures for measuring mass and energy, and procedures for predicting and verifying observable behavior of the thing under specified verifiable conditions. In essence, modern physics defines *physical* in observation terms in contrast to non-observable metaphysical terms.

For the things we ordinarily consider physical, there are observation and measurement procedures for verifying that they are in fact physical. In any case, the final step is the conscious awareness of a pointer reading, a digital display, a beep or flash, etc. But what is the method for deciding whether or not a conscious experience is physical? While we sometimes talk as if ideas have weight and energy, these are only metaphors. We have no scientific procedure to directly measure the mass and/or energy of conscious experiences. Since conscious experience is inherently private, it is not clear how one could ever verify that another person had a specific conscious experience, let alone whether it is physical. On the other hand, like Descartes, I have no doubt that I have conscious experiences.[2]

A physicalist typically argues that associated with every conscious experience is a specific physical neural process in that

[2] This assertion raises the question of what is the referent of the pronoun "I", and whether the referent is physical or mental. This issue will be taken up in Chapter IV.

experiencer's physical brain, so a specific conscious experience could in principle be verified by measuring neural activity. However, this assertion does not resolve the problem, since it relies on the assumption that whenever a specific neural activity occurs the associated conscious experience necessarily occurs. It is an assumption because the occurrence of the conscious experience itself is unverifiable. Granted open brain surgery on a conscious patient has demonstrated that stimulation of specific locations in the brain produce verbal reports of a conscious experience. Nonetheless, there is no way to independently verify the validity of these reports since the reports in principle could be caused by the neural activity without there being any conscious experience: a.k.a. *the zombie problem*.

Even if there is a mapping from neuro-states into conscious experiences, it does not follow that a conscious experience is *identical* to the co-existing physical neuro-states. Indeed, the conscious experience has qualities like color and pain but not mass and energy, while the physical state has mass and energy but not color or pain. To say that they are identical is a misuse of the word "identical".

Lightning is sometimes erroneously given as an illustration of identity. As Ben Franklin demonstrated, lightning is associated with a sudden flow of electricity from clouds to earth (or clouds). But is it legitimate to say that "lightning is identical to a sudden flow of electricity from clouds to earth"? The answer depends on

what the word "lightning" refers to. Prior to Franklin's discovery, the underlying cause of lightning was unknown, so the referent of lightning could have been "the anger of the thunder god", or "whatever physical process causes the flash of light and sound in the atmosphere". For clarity, let **lightning$_0$** denote this latter meaning. In this case, after Franklin, we could say that **lightning$_0$** is identical to the sudden flow of electricity from clouds to earth. On the other hand, suppose "lightning" refers to the *conscious experience* of a bolt of light in the atmosphere, and for clarity let **lightning$_1$** denote this meaning. Since a sudden flow of electricity from clouds to earth does <u>not</u> explain the conscious experience, **lightning$_1$** is not identical to a sudden flow of electricity from clouds to earth. At most, one could conclude that lightning$_1$ implies there was a sudden flow of electricity from clouds to earth. [The converse does not follow because there could be many instances of sudden flows of electricity from the clouds to earth that are not consciously experienced as lightning$_1$ by any human.] Therefore, the statement that "the conscious experience of lightning$_1$ is *identical* to a sudden flow of electricity from the clouds to earth" is categorically wrong, just as the statement that "a conscious experience is *identical* to the co-existing physical neuro-states" is a categorical mistake.

In epistemology, there is a distinction between empirical knowledge and analytic knowledge. The former is about the physical world. The latter includes logic and mathematics. The truth of analytical statements does not depend on the physical

world. Furthermore, analytical truths have no mass or energy as defined by current science. That is, analytical truths are non-physical abstractions. However, a physicalist might say that to comprehend an analytical truth requires a physical brain. Does an analytical truth (e.g. a simple syllogism) "exist" if not comprehended? To assert that it does not exist takes us down a dangerous path. There are many true mathematical theorems that I do not comprehend but are comprehended by some mathematicians. Does the mathematical theorem exist for this group of mathematicians but not for me? How strange, since then existence is relative to individual brains. The physicalist would never apply this line of reasoning to physical theories, for then the existence of electromagnetic forces (etc.) would be relative to individual brains.

I am quite willing to accept that some mathematicians have established the truth of a theorem and comprehend it, and therefore I am willing to believe that this theorem is true (i.e. exists as a true mathematical theorem) even though I cannot comprehend it, just as I am willing to believe in Einstein's general relativity even though I do not comprehend the equations which define it. Further, I am willing to believe that these theorems and theories are *timeless*.[3] It seems that the physicalist wants me to believe that at the time Pythagoras first formulated the famous Pythagorean Theorem, but

[3] Since an abstraction is non-physical, it does not exist in physical space or time, and so any statement about when a theorem became true (such as before or after humans appeared) entails a categorical error.

before he had a proof, the Theorem was neither true nor false. While I agree that we did not know whether the Theorem was true or false until a proof was found and verified, I have to believe that it was true all along.

A resolution of this problem can be attained by distinguishing another meaning of "exist". Since the truth of an analytical statement does not depend on the physical world, although the statement could be instantiated in many physical ways (e.g. written or spoken in various languages), the thing that is common to all these instantiations is an abstract idea. Hence, one should say that analytical truths *exist as abstraction.* Moreover, one can believe there are true analytical truths that have not yet been discovered by any human. It is not clear the physicalist would agree, thereby admitting to the "existence" of non-physical things. Instead he may simply deny the usefulness of the concept of *existence-as-an-abstraction,* but then the physicalist must conclude that mathematics is useless.

There is also the physicalist argument from analogy with vitalism. From Aristotle to the 19th century, it was argued that a purely physical description of a living organism cannot possibly explain the ineffable quality of being alive. Nonetheless, as biochemistry advanced, more and more observable aspects of being alive were described in terms of chemical reactions, until vitalism joined the ranks of geocentrism and other discarded ideas. Similarly it is hoped by some physicalists that someday there will

be a purely physical explanation of conscious experience. However, the analogy is flawed because vitalism was replaced by ever more detailed physical explanations for *observed behavior* previously thought to be unexplainable by physical processes. In contrast, private conscious experience is not an observable behavior and so there will remain a gap between the ever more detailed physical processes and the private conscious experiences.

In conclusion, physicalism appears to be refuted by the simple observation that I have private conscious experiences and they are not physical as defined by current science. Of course, the physicalist could assert that the current science definition is inadequate and should be expanded to include ideas and conscious experience.[4] However, that route would merely render the statement that `everything is physical' a tautology. Moreover treating "physical" as the name of the super class of all things does not solve the hard problem of how one subclass of "physical" (neuro-processes) generates another subclass of "physical" (conscious experiences).

C. Mentalism (Idealism).

One of the earliest philosophical arguments for the existence of non-physical abstract things (e.g. ideas and concepts) lies in the writings of Plato (375 BC). For example, take the concept of a

[4] Panpsychism takes this approach, but I seriously doubt that real physicists (as opposed to metaphysicists) will ever accept an unverifiable property as a fundamental characteristic of the physical world.

perfect circle: there are many imperfect circular physical objects, but no perfectly circular physical object. Thus, the concept of a perfect circle cannot be a physical thing; rather it is an abstract non-physical thing. Clearly the concept of a perfect circle exists as an abstraction; therefore non-physical things "exist" in some sense. Hence, the word "exist" has at least two very different meanings: (i) a rock *exists as a physical object*, and (ii) a perfect circle *exists as an abstraction*. To ask where abstractions reside is to misunderstand the difference in the two meanings. Since abstractions are not physical, they do not reside in physical space-time.

Plato argued that abstract forms are more fundamental than physical objects. Mathematicians naturally embrace this Platonic mentalism, since mathematics entails thinking about abstract things (numbers, spaces, sets, and logical operators). A true mathematical/logical theorem *exists as an abstraction* and does not depend on whether the proof is spoken, written on paper or stored digitally.

Science has supplemented our senses with precise instruments to measure properties of physical things. However, all these advancements have not removed the requirement of a conscious experience (e.g. seeing a blip on a screen, hearing a click, or reading a number on a display) as a necessary step in the verification process. As argued above, the physicalist's assertion that physical neuro-processes of the brain when seeing a blip on a

screen etc. is identical to the conscious experience is a categorical mistake. However, it does not follow that one must be a dualist.

Berkeley (1710) is usually credited as the first philosopher to seriously deny the existence of physical things. The ancients were aware of the problem of optical and auditory illusions and therefore distinguished between the perception of a thing and the thing in-itself. Descartes followed the skeptical path to the conclusion that all he could be sure of was the existence of his thoughts, but he stopped short of actually denying the existence of physical things. In contrast, Berkeley asserted that since his perception of "reality" consists solely of conscious experiences, the belief in an objective physical reality is an illusion. To be untethered from an objective reality is a scary state of affairs, and Berkeley found serenity in the belief that God created his conscious experiences.

D. Phenomenism.

Kant (1781), while not denying the existence of things-in-themselves, asserted that we can never know anything about them. The notion of an impenetrable veil between us and things-in-themselves can be traced back to Plato's metaphor of "shadows on a cave wall" [*Republic VII*]. All the things we perceive are *phenomena*: that is, conscious experiences or mental things constructed from conscious experiences. Therefore, the only things we can sure of are (i) analytic truths, and (ii) our conscious

experiences (as conscious experiences)[5]. In contrast to Berkeley who invoked a supernatural being that guides phenomenal reality, Kant assumed there is an objective but unknowable *noumenal* reality that underlies phenomenal reality. This assumption distinguishes phenomenism from Berkley's mentalism.

Phenomenism has undergone considerable change since Kant. J. S. Mill (1843) argued that physical objects do not cease to exist when not perceived because they stand for "permanent possibilities of sensation" (whatever that means). As science discovered more about the physics of perception, another notion was that physical objects are "bundles of sense-data", where *sense-data"* denote the encoding of physical inputs to the body (light, sound, etc.) into mental states anterior to conscious perception. Ernst Mach (1883) resisted this notion and considered conscious experience to be the raw data. In contrast, the logical positivists embraced the concept of sense-data and embarked on an effort to construct a theory of phenomenal reality in which sense-data were the fundamental elements. In their view, all statements about phenomenal objects could be translated into statements about only sense-data. This effort is now considered as having failed due mainly to not

[5] The parenthetical is added to emphasize that a perception does not imply the existence of what the conscious experience appears to be. That is, the conscious experience of a red ball does not imply the existence of a physical red ball – only that I am having a 'red ball like' conscious experience. It might be that I am being deceived by a magician, or looking at a white ball illuminated by red light.

recognizing the necessity of a larger context (theory/model) to give meaning to statements about phenomena [Sellars, 1963].

Twentieth century scientific theories of perception, by taking things-in-themselves as given and deriving how they are perceived, implicitly presume that things-in-themselves are knowable – contrary to Kant's view. Once noumenal reality is regarded as unknowable, further deliberation about things-in-themselves is a meaningless waste of time.

Following Kant, modern day phenomenism holds that we can never truly know anything about objective (a.k.a. noumenal) reality - all we can perceive and know about are *phenomena*. Nonetheless, it is natural to wonder what is out there causing our experiences. Fortunately, we do not have to start from scratch, but instead we are endowed with a brain that has genetically coded models of what's out there. Neuroscience has discovered many of the neural mechanisms that take my sensory data and construct a model of my body and my local environment. Physics has provided us with models of physical things that has proven reliable and accurate.

E. Summary.

Dualism is plagued by the problems of the interaction between the physical and mental, and why mental things are not superfluous. Physicalism is refuted by the fact that (i) abstract (non-physical) concepts, such as perfect circles and mathematics,

play an irreplaceable role in science, and (ii) the unfounded assumption that conscious experiences and neural processes are identical. Mentalism, al la Berkeley, is incomplete without blind faith in a supernatural creator and guider, and even then it provides no reliable and accurate method of predicting future mental states. Phenomenism avoids the foregoing problems, but there is still a need for more clarification of the relationship between conscious experiences and phenomenal reality.

III. Models, Conscious Experience and Phenomenal Reality.

I agree with the common insight of Descartes (1641), Hume (1748) and Kant (1781) that (i) we can never know anything about *noumenal* reality[6], and (ii) the only non-analytic things I can know for sure are my conscious experiences (as conscious experiences)[7]. Kant called the things of conscience experience *phenomena* to distinguish them from the unknowable things-in-themselves of

[6] The notion of an impenetrable veil between us and things-in-themselves can be traced back to Plato's metaphor of "shadows on a cave wall" [*Republic VII*].

[7] The parenthetical is added to emphasize that a conscious experience does not imply the existence of what the conscious experience appears to be. That is, the conscious experience of a red ball does not imply the existence of a physical (noumenal) red ball – only that I am having a 'red ball like' conscious experience; it might be that I am being deceived by a magician, or looking at a white ball illuminated by red light.

noumenal reality. Phenomena are abstract objects (such as trees and birds) and relationships between those objects (such as distance and motion).[8] Hence, the things of conscience experience are abstract objects and relationships between those objects.

A term for a system of abstract objects and relationships between those objects is a *model*. Thus, phenomenal reality can be viewed as a collection of models. I am partial to this view because it strongly conveys the distinction between phenomenal reality and noumenal reality. Further, I can state that ***a conscious experience is the perception of a model of phenomenal reality.***

Since the concept of a model is central to what I am trying to convey, I will begin in Section A with a definition of "model" as I intend it to be interpreted. The concept of my *here&now* model will be defined as well as my *background* models. With this foundation, Section B will delve into what conscious experiences are, distinguishing conscious sensual experiences and conscious thought experiences. Section C tackles the question of what comes first, a model or a conscious experience. Section D addresses the *hard problem of consciousness*. Finally, Section E concludes that my *phenomenal reality* is the collection of my here&now model

[8] Note that in these examples I reinterpret the "physicalist" terms (tree and distance) as non-physical abstractions. Indeed, Kantian phenomenalism demands that we reinterpret physicalist terms as non-physical abstractions or else invent cumbersome new terms such as "ϕ-tree" and "ϕ-distance".

and my background models. In addition, it sets forth the premises of Scientific Phenomenism.

A. Models.

Formally, **a model is a collection of objects, relationships between those objects, and a law of motion that determines how these relationships change over time**. A model is an abstract (mental) thing transcending the medium in which it is presented (such as on paper, in digital bits, or in neural patterns). Further, the objects in a model are abstract things even though they may be called rocks or dogs. In other words, the things in a model are **phenomena**, not noumenal things-in-themselves. Unfortunately, everyday English allows us to talk of a "model of X", which can mislead us into thinking that X is undeniably real (i.e. a noumenal thing-in-itself) and that the model "represents" X. Clearly, this way of thinking is incompatible with phenomenism. Instead, we should interpret "of X" as merely indicating a specific model in the class of all models: e.g. an "X-model".

This notion of model contrasts sharply with a child's toy model train or an architect's model of a skyscraper. The objects in those kinds of models are typically physical representations of much larger physical objects. A possible synonym for "model" in the sense I use the term could be "theory" but a model has much more detail than a general theory.

The Standard Model of particle physics is a model in the sense I use the term. As just discussed, "of particle physics" does not imply that particle physics is part of noumenal reality, but merely indicates which "Standard Model" in the class of all models. It is an abstract entity consisting of a collection of abstract objects (symbols for particles), their relationships with each other, and a dynamic of change. The relationships and dynamics are defined explicitly by logic and mathematics. Nevertheless, some details are left out (e.g. gravity and dark energy). Cosmologists study models of the universe (i.e. universe-models), but those models do not contain the details about every particle; often whole galaxies are treated as homogenous objects. Similarly, ecologists study models in which plants of a kind are homogenous objects; biologists study models in which molecules of a specific kind are homogenous objects; chemists study models in which atoms of a specific kind are homogeneous objects; and physicists study models in which the fundamental particles are homogeneous objects. Economists have models in which consumers have utility functions, producers have production functions, and trade takes place in markets with uniform prices.

Often the objects left out of a model are of the same kind as the objects in the model. For example, we have models of weather on Earth that leave out the other planets in our solar system. Since the orbits of other planets do affect Earth, such a simplified model cannot possibly account for all the observables about Earth. On the other hand, we have models of our solar system that account

for the orbits of all planets. If we embed a model of weather on Earth into the model of the solar system, we will have a model in which the weather on Earth is affected by the other planets. Typically, in this expanded model we divide the variables into two sets: (i) those which pertain to observables about Earth (called endogenous), and (ii) those that pertain to the other planets (called exogenous). As it turns out, the effect of the other planets is miniscule in comparison to the effect of Earth's moon, so an earth-moon model (taking the exogenous variables as constants) will suffice for the pragmatic purpose of predicting weather on Earth. From the perspective of the earth-moon model, the exogenous variables are simply given and not explained, but from the perspective of the solar-system model, those variables become endogenous to the solar-system model and thereby their specific values are explained. In other words, a model explains the values of its endogenous variables but does not explain the values of its exogenous variables. Note that "earth" can refer to either (i) an object in the solar-system model or (ii) a model of Earth containing the water and mountains, the mantle, the liquid iron core, etc.

Neuroscience has discovered many of the neural mechanisms that take sensory inputs and construct a model of my local environment (e.g. Ulanovsky, 2011). For ease of reference, let me call this model my ***here&now*** *model*. Objects in this model are located by three distance coordinates and one time coordinate relative to me and now. What is "local" changes in spatial perspective and focus: e.g., when I am looking into a microscope,

when I am typing on a computer, when I am driving a car, and when I am gazing at the night sky through a telescope.

As I slowly turn my head, the images on my retina change and my here&now model changes, but generally I do not perceive my surroundings as moving and my head as being stationary. Instead, I generally perceive my surroundings as being spatially fixed and my head as moving. Apparently, from the time varying here&now model, my brain constructs a (subconscious) *background model* that has a spatial and temporal scope larger than my here&now model, and this background model is the basis of my expectation of what my here&now model will be like as I move my head and body. For example, if I am looking at a coffee mug on my desk and I make a 360 degree turn, I expect to see the same coffee mug in the same location at the completion of the turn, because in my brain's background model there is an object standing for the coffee mug that exists at a fixed location in that model even when I turn 180 degrees away and cannot see it. In other words, my brain's background model incorporates *object permanence*. Of course, I could have been wrong. Perhaps a magician arranged a mirror so an image of a mug appeared, but when I turned away the magician removed the mirror. However, experiences such as not seeing a mug after making a 360 degree turn have been so infrequent that rather than discarding object permanence, I call up an alternative background model (such as one with a magician) that is compatible with my experience and object permanence. Should there be a

significant inconsistency between my expectation and my background model, my brain raises an alarm.

In addition, I have models about hypothetical/imagined worlds. For instance, Euclidean geometry is a model of an imagined world that obeys the axioms of Euclidean geometry. I also have models of physics, chemistry, biology, and psychology. I will call these my *background models*, In contrast to my here&now model, my background models about hypothetical/imagined worlds are not firmly tied to my current sensory inputs. Since my background model is created from my time-varying here&now model, my background model is obviously linked to my sensory inputs, but being an extrapolation, it is technically hypothetical. Humans would not have developed our current technology without the ability to create models of hypothetical/imagined worlds, to think about them, and to judge which provide more accurate expectations/predictions. Indeed, this essay is such an exercise in creating and analyzing a hypothetical/imagined model.

In summary, the term "model" entails three important features.

(i) It preserves the distinction between the unknowable (noumenal reality) and the knowable phenomenal reality; in particular, a model is knowable.

(ii) A model is a wholistic concept in contrast to sense-data theory which is bottom-up.[9] By being a whole construct no individual object in a model has a meaning by itself, but only in relation to other objects in the model.[10]

(iii) Objects in a model can have permanence without requiring permanence of perception.

B. Conscious Experiences.

Since phenomenism asserts that all the non-analytic things we perceive are conscious experiences or abstract things constructed from conscious experiences, we need to delve more deeply into the nature of conscious experiences. First, I want to make a clear distinction between *awareness* and *consciousness*. Awareness entails merely responsiveness or a disposition to respond without

[9] According to sense-data theory, all empirical sentences are translatable into sentences about sense-data which are the building blocks of perception. It turns out that this task is impossible without reference to relational laws. In Kuhn's (1962) terms, observations (i.e. phenomenal objects) are "theory-laden."

[10] E.g. an electron, defined as a fundamental particle with a radius of 10^{-22} meters, a mass of 9.1×10^{-28} grams and an electric charge of -1.6×10^{-19} coulombs, has meaning only in reference to an electromagnetic field which in turn depends on the spatial distribution of all other charged particles. We can say nothing about the behavior of an electron without specifying the electromagnetic field in its vicinity. Moreover, we also need at least four other particles with respect to which we can measure distance and direction in 3D space. In other words, an electron is an object in a model that contains various kinds of objects (such as protons) and the relationships between those objects, and laws of motion; by itself an electron has no meaning. A model can be analyzed in terms of its components (such as electrons), but in general, the components cannot be separated from the whole model without losing meaning.

any cognition about what one is aware of or doing. For example, reflex reactions and autonomic behaviors imply awareness but not cognition. In contrast, *consciousness entails cognition of a model of the world*.

My conscious experiences are profoundly private and inaccessible by anyone besides me. In other words, you cannot know or deduce my conscious experiences. Similarly, while I have direct knowledge of my conscious experiences, I cannot know or deduce that you would have the same or similar conscious experiences in identical situations. Consequently, this essay can only be written from my 1st-person perspective.

My stream of consciousness is a sequence of sensual experiences and also thought experiences. As I look out my window now, my experience is primarily sensual, specifically visual 3D images in various colors and lightness. As I pause to type these words, my experience changes to primarily thoughts about sentence structure and spelling, while my visual experience beyond my window fades in consciousness. In other words, there are essentially two fundamental categories of conscious experience. The first category is *sensual* and the objects in the model can be called "physical" because they relate to each other according to the folk laws of physics. The second category is *thought* and the objects in the model are abstract (non-physical) and related to each other by definition, logic and mathematics. Rather than physical objects and abstract objects being different

substances (as in dualism), the adjectives physical and abstract refer to the different kinds of relationships in a model.

1. Conscious Sensual Experiences.

My conscious sensual experience is my perception[11] of what is happening here and now. It consists of *objects* such as rocks, trees, rivers, animals, buildings, cars, clouds, etc., and *relationships* between those objects such as distance and direction (in 3D space relative to my self), bigger than, lighter than, redder than, louder than, sweeter than, more pungent than, smoother than, earlier than, faster than, etc. Together, these objects and relationships constitute a model of my here and now world, which I call my ***here&now model***. When I describe my conscious sensual experiences, my statements are about this here&now model.[12] My here&now model interprets my sensual experiences, and to the extent that those sensual experiences exhibit regularities, the relationships in my here&now model will also exhibit regularities: such as physical objects at rest will remain at rest unless acted upon by a physical force. In other words, the "physical" objects will be related to other "physical" objects in "physical" ways. Note, however, that these "physical" objects are still *phenomena*

[11] Perception implies a perceiver. Obviously, the perceiver is my self. But what exactly is my self? Chapter IV will delve into this question and provide a non-dualist answer.

[12] A statistician would call my here&now model a Data Generating Process (DGP), where the data are sensory experiences. Just as DGPs provide interpretations of the data for the statistician, my here&now model provides interpretations of my sensory data.

rather than noumenal things-in-themselves, so a mind-body problem does not arise.

In contrast, the neuroscience model tells a temporal story of the cones in my retina being activated first, and then nerve impulses being generated and flowing to other neurons that respond to edges and shapes and eventually activating neurons in my visual cortex. However, I do not experience this temporal sequence of neural activities as a temporal sequence, but instead I experience green trees, blue flowers and brown rocks as objects in my here&now model. Therefore, the neuroscience model does not explain conscious sensual experience.

To further elucidate the concept of my here&now model, consider the following observations.

- The objects in my here&now model can be stationary or moving in 3D space (e.g. the bird that just flew by my window). The fact that I do experience objects moving implies that my experience event has a non-zero temporal width, so I can perceive that an object has a different location at the end of the event than at the beginning of the event. This detection of movement can also produce a quantitative assessment of velocity. While I may have only a vague sense of this quantification, I do experience surprise when in the next event some object appears at a location that is inconsistent with it traveling at the velocity

it had in the previous event. In other words, I experience movement and changes in velocity (i.e. acceleration).

- Just as I don't typically view a digital picture through a magnifying glass with a field of vision restricted to a single pixel, my conscious visual experience does not typically have a field of vision so microscopic that it appears as a homogeneous patch of light. Neither does my conscious visual experience consist of an enormously large unorganized array of patches of light. Rather, it consists of distinguishable objects and relationships between them. Neuroscience has made progress in discovering neural mechanisms that are involved in transforming the incidence of light on my retina into a here&now model.

- Rapid eye movements (saccades) produce a sequence of different images on my retina, but I am not conscious of these distinct images; instead, I have a stable visual experience of the whole area scanned by the saccades. Hence, my here&now model covers an area larger than any one retinal image. It has a focal point with clarity and detail highest at the focal point and decreasing towards the periphery. Holding my head still, the extent that clarity and detail decrease with distance from the focal point depends on my level of attention. For example, as I am typing these words, my attention is concentrated at the

cursor on my laptop screen, and clarity and detail fall rapidly with distance from the cursor. However, when I stop typing, my attention and the clarity and detail of my visual experience spread out. That is, clarity and detail increase towards the periphery of my field of vision as it diminishes at the focal point. The neuroscience model of vision explains this variable attention as due to variable weights on the inputs to the visual neural network from the rods and cones of the retina.

- As I sit here at my laptop and turn my head, the content of my field of vision changes as it sweeps over areas that were peripheral, but instead of perceiving my room as moving, I perceive it as stationary and my head as moving. When I return my head to my laptop screen and my field of vision sweeps over the same areas in reverse order, I not surprised when I see my laptop again because my brain has transformed sensual inputs into the objects and relationships of my here&now model, and it has constructed and stored a stable background model from past as well as current sensual inputs; hence I perceive my laptop as stationary and my head as moving. In other words, the stable background model constructed by my brain has object permanence, and my conscious experience incorporates this interpretation.

- It is reasonable to suppose that I have stored in memory background models for general categories of situations (such as typing on my laptop, eating in a restaurant, conversing with my wife, hiking a mountain trail, etc.), and variations of these models for subcategories. These models adapt over time in response to conscious sensory experiences. In addition, past sensual experiences of using a street map or reading about science could also influence these background models, thereby influencing my current here&now model. Moreover, these models are ready to be activated by current sensory inputs to provide context for those experiences and to initiate reactions by me.

- One of the important objects in my here&now model is my body. The object standing for my body (my *body-object*) can come in many versions with different levels of detail. For example, when I am gazing into the distance from a 14,000 ft mountain peak, my body-object may have few details other than my head and eyes. As I am writing now, my body-object also has hands and fingers. When I focus my attention on my body, my body-object has a brain, a heart, and other anatomical features.

- Recalled memories of sensual experiences can be considered a subclass of conscious sensual experiences. However, the details and the intensity of the sensual

qualities of a recalled memory can be considerably attenuated.

Importantly, since my conscious experiences are inaccessible to anyone but me, the sensual qualities (a.k.a. qualia) of my conscious experiences are also inaccessible to you and hence irrelevant to you. For example, while my color experience of a **red** pen is "red",[13] your color experience of that same pen could be more like my color experience of a **blue** pen (i.e. "blue"). To understand how this could happen suppose that (i) when my red cones are excited by a **red** pen, those cones induce a cascade of neural activity that I experience as "red"; however, (ii) when your red cones are excited by the same pen, those cones induce a cascade of neural activity that is exactly like the neural activity I have when I see a **blue** pen, so you experience "blue". It just so happened that from childhood onward, you have learned to use the word red when referring to your conscious experience of "blue". Moreover, just as my brain has recorded a correlation between seeing "red" and emotional anxiety, your brain will have recorded a correlation between seeing "blue" and emotional anxiety. Because your conscious experience is inaccessible to me, I cannot logically or empirically refute this possibility. On the other hand, when you say you see a **red** pen, I can infer that if I look at the same pen, I

[13] I am using quotes around a color to indicate that this color refers to the sensual quality of my conscious experience rather than the objective wavelength (700 nanometers) of the light incident on my retina, and I am using boldface to indicate the objective wavelength.

will see a **red** pen and have the same color and emotional experience that I always have when seeing a **red** pen. Moreover, this inference is verifiable. Thus, given we have learned a common language, your statements about your color experience transmit practical information to me about what I would experience in the same situation, but your statements convey no information about the private qualia of your conscious experience.

Does the above argument apply to other senses such as hot and cold? Imagine that your hot and cold temperature sensors induce the reverse neural cascade as mine. When I enter a **hot** sauna, I begin to sweat and feel "hot". When you enter a **hot** sauna, you also begin to sweat but you feel "cold". Nonetheless, you have learned to use the word hot to refer to your "cold" sensation. Hence, you would say that the hot sauna causes your body to sweat. Conversely, when you enter **cold** water, you begin to shiver and feel "hot", but you would say that the cold water causes your body to shiver. Moreover, hearing your testament I can be confident that if I entered the same **cold** water, I would feel "cold" and begin to shiver like you. As with colors, your statements about hot and cold convey no information about the private qualia of your conscious experience. A similar argument can be applied to the other three senses.

Similarly, the qualia of my sensual experiences are totally hidden from you and irrelevant to you. Therefore, qualia are absent from the 3rd-person perspective of science. However, *from*

my 1st-person perspective qualia are inseparable undeniable aspects of my conscious sensual experiences.

2. Conscious Thought Experiences.

My conscious thought experience is the contemplation of a *virtual* world not necessarily constrained by sensual inputs, and takes the form of a model. For example, when I am thinking about whether it might snow tomorrow, I am contemplating a model of future weather. When I am trying to remember what I had for dinner last night, I am contemplating a reconstruction of a past here&now model. When I am wondering why there is a post-pandemic shortage of labor, I am contemplating a model of the economy.

Much of the time, I am thinking about abstract things: philosophy, mathematics, physics, and psychology. These conscious experiences take the form of *models*: abstract objects and relationships between those objects. I feel as if I am in a space different from the space of my sensual experiences. Mathematics has given me the ability to conceptualize abstract spaces of more than three dimensions and unlike the Euclidean space we normally use as the framework for our model of sensual experiences. Therefore, I have no problem conceptualizing an abstract space as a framework for thinking (or symbol manipulation). If asked "where are my thoughts located", I am not embarrassed to say that they exist in an abstract space of thinking not the 3D space of sensual experience.

Much of my thinking takes the form of dialogues in words. Words themselves are abstract objects. The word "rock" has the same meaning whether written in Courier or Times Roman font, spoken softly or loudly, or expressed in French; hence, a word is an abstract object independent of its experienced form. Words are combined into sentences which narrate an aspect of a model. Nouns refer to objects in a model and verbs refer to relationships. *Sentences have meaning solely in terms of a model.*

I have many virtual-world models: (i) some with narrow scope (e.g. the room I am sitting in now) that are heavily influenced by my recent here&now models; (ii) some with medium scope (e.g. my neighborhood) that are influenced by a mix of past here&now models and external models such as street maps; and (iii) some with broad scope (e.g. the solar system) that are heavily influenced by models provided by science. While science-based models are internally consistent, many of my virtual-world models are not internally consistent, and finding an inconsistency motivates me to fix my model or reject it for an alternative model.

A categorical difference between any of my models (here&now or virtual-world) and science-based models is that the former are internal to me while the latter exist in many formats (written or digital) external to me. For example, neuroscience has discovered many of the neural mechanisms that take sensory inputs and construct components of my here&now model [e.g. Barrett, 2021]. This neuroscience model is external and accessible to me and you,

while my here&now model is internal to me and inaccessible to you.

Science also has models within models; e.g. a chemical model whose objects (molecules) are comprised of atoms, and an atomic model whose objects (atoms) are comprised of electrons, protons and neutrons, etc. Similarly, I have less formal models within models; e.g. a neighborhood model whose objects (families) are comprised of people, and a person model whose objects are bodies.

It is reasonable to suppose that my brain as stored *templates* of virtual-world models for general categories of situations (such as typing on my laptop, eating in a restaurant, conversing with my wife, hiking a mountain trail, etc.), and variations of these models for subcategories. These templates are ready to be activated by current sensory experiences to provide context for those experiences. Further, these templates adapt over time in response to my stream of sensory experiences.

I feel the need to caution the reader to not interpret objects in any model as *representations* of noumenal things-in-themselves. To say that a phenomenal object is a representation of some thing-in-itself is non-sensical in phenomenism because things-in-themselves are unknowable. In other words, the assertion that X is a representation of Y cannot be verified as true or false when Y is unknowable, so the assertion is vacuous. On the other hand, an abstract model can be a representation of a phenomenal model because both are phenomena.

C. What comes first, a model or a conscious experience?

Since phenomenism asserts that the only non-analytic things we can be sure of are conscious experiences, it might seem that conscious experiences come first, then the models constructed from a history of conscious experiences. But I do not have any evidence that I had conscious experiences at the moment of birth. However, at birth I had a brain that, through millions of years of evolution, was endowed with many innate models of the world I was entering. As my sensory organs and brain matured, my sensory stream activated these innate models which provided a meaningful interpretation of that sensory stream. Thus, I cannot disprove that my models came first, then my conscious experiences interpreted by those models. As I matured, I learned refinements of these models as well as new models (such as classical physics). I also learned/imagined abstract models whose usefulness could be confronted and tested by my conscious experiences. By utilizing the scientific method, my storehouse of useful models greatly expanded.

D. The Hard Problem of Consciousness.

A very important difference between my here&now model and a science-based model is that the former entails a 1st-person

perspective while the latter entails a 3rd-person perspective. For example, as I look at the pen on my desk, from my 1st-person perspective (my here&now model), the pen has a red sensual quality (i.e. appears red to me); but this sentence, the word "red", and the 3rd-person neuroscience model of my looking at the pen contain no red sensual qualities whatsoever. This observation raises the so-called *hard problem of consciousness* [Chalmers, 2002]. That is, so far we have failed to construct a science-based model that entails conscious experiences as essential elements and not just as epiphenomena.[14] However, since my conscious experiences are inaccessible to anyone but me, no 3rd-person science-based model can provide an essential role for the sensual qualities of my 1st-person here&now model. In other words, the hard problem of consciousness (as posed) is unsolvable, and therefore we should waste no more time trying to solve it.

In contrast, phenomenism takes the existence of conscious experience and its qualities as fundamental/given. There is no need to construct a science-based model that entails conscious experiences as *implied* elements. Nor do we fall into denial of noumenal reality (as did Berkeley). Rather, the pertinent problem of consciousness is: "how it is possible that from conscious experiences we can discover models of phenomenal reality that are reliable and useful?" The answer is that by using the scientific method, which demands the testing of the predictions of

[14] For further elaboration of this point refer back to my critique of physicalism in Chapter II.B.

hypothetical models, and by applying statistical analysis, we can quantify the reliability of these models. Furthermore, our evolutionary survival depends on using the current most reliable models when making decisions. An informative name for this point of view is ***Scientific Phenomenism***.

An open issue is the need for a model that gives meaning to the concept of a 1st-person (a self) as distinct from a 3rd-person. This issue will be addressed in Chapter IV on "My Self and Models of My Self".

E. Phenomenal Reality and Scientific Phenomenism

In summary, a conscious experience is the perception of a model that interprets and gives meaning to the experience. For conscious sensual experiences, the model is my here&now model. For conscious thought experiences, the model is about virtual worlds. At any moment, one model is activated, while the others are stored in memory ready to be activated when triggered by sensual inputs. A model is a wholistic concept in which objects are distinctive by virtue of their relationships with all other objects in the model, in contrast to a reductionist bottom-up concept. Many of the objects in my models are "physical" by virtue of their being related to other objects in "physical" ways. Other objects are "non-physical" by virtue of being related to other objects in non-physical ways, such as definitional, logical or mathematical.

Nonetheless, all objects and relationships are *phenomena* as opposed to noumenal things-in-themselves.

So, given all these models and phenomena, what is *phenomenal reality*? The concept of phenomenal reality is not well-defined in our everyday language or in philosophy. It could denote a conscious experience or a stream of conscious experiences. However, to be some kind of "reality" I would not include conscious thought experiences of virtual worlds. Still, allowing only my here&now model would leave out my stable background models. Since these background models also consist of phenomena (objects and relationships) and constitute hypotheses about my world, these background models should be included.

Therefore, I define **my phenomenal reality** to be **the collection of my current here&now model and my background models.** Note that as a collection of models, my phenomenal reality is an abstract object like the models and objects it entails.

Since these models are not necessarily consistent with each other, my phenomenal reality is likely to have many inconsistencies. While perhaps similar to your phenomenal reality, my phenomenal reality is not the same as yours. I have a unique private perspective, a unique history of conscious experiences, and unique models that interpret those experiences.

An informative name of the view I have developed here is ***Scientific Phenomenism***. It consists of several premises. First,

the only non-analytic things I can be sure of are my private conscious experiences. Second, evolution has provided me with a brain that (i) maps sensory inputs into models of my world that interpret those inputs, and (ii) makes statements about these models. These models are abstract, as are the objects and relationships in these models. Third, my phenomenal reality is the collection of my current here&now model and my background models. Fourth, by using the scientific method, which demands the testing of the predictions of hypothetical models, and by applying statistical analysis, we can quantify the reliability of these models, reach some consensus regarding reliable models, thereby improving our chances of survival.

IV. My Self and Models of My Self

A. Introduction.

We use the first-person singular pronoun "I" in many ways: I see, I hear, I speak, I walk, I think, I believe, etc. What is this "I" that is doing these things, having these experiences? A common answer is that it is my ***self***. It is important to emphasize that in this usage of the pronoun "my", the possessive interpretation is unintended and inappropriate. That is, "my" does not imply the existence of some other entity that possesses or owns that self (as in "my shirt"), but rather it is merely a pointer to **that** self, in the class of all selves, to which "I" refers.

But what exactly is my self? A Google search will yield many incompatible notions of self including physical (e.g. my brain), mental (e.g. my mind/soul), and an illusion. Each notion entails criticisms of other notions, and there is obviously insufficient space in this chapter to present an enlightening

review. The modern literature is reviewed by Gallagher and Sheer (2000) and Zahavi (2008).

The lack of a consensus answer to this question is curious. A grammatically similar question such as 'what is an automobile?' immediately conjures appropriate answers such as 'a four-wheeled gasoline or electric powered vehicle to transport people from one place to another'. Implicit in this answer is an assumed context: a world with roads connecting distant locations and various means of transporting people between those locations. In other words, there is a *model of the world* that contains objects like people, cities, houses, roads, bicycles and automobiles. Indeed, the answer to `what is an automobile' has meaning only in reference to such a model. A realist would object, and assert instead that the referents of our nouns are the real objects in the real world. However, this realist view is not compatible with well-known visual illusions and with virtual reality googles in which virtual objects feel real and outside our body in 3D space.

The question 'what is my self' is more difficult to answer because it is not always clear what object in what model corresponds to my self (e.g. see Gallagher and Sheer, 2000; and Metzinger, 2007). Since the concept of a *model* is central to the view I am developing in this essay, I refer the reader back to Chapter III.A. A fundamental premise is that **meaningful sentences - whether written, spoken or thought – are about a model and have meaning for the author/speaker/thinker only**

in reference to that model. Applying this premise to sentences involving "I" will reveal that the referent of "I" is often different kinds of objects in different kinds of models.

Section B examines sentences in which the referent of "I" is merely my body (which I call my "proto-self"). When speaking of the internal structure of my body, my proto-self is a model of my body. When speaking of my body and objects external to my body, my proto-self is an object in a model of my world. I formally define a *1st-order model* of my world as a model with my proto-self and objects external to my proto-self.

In Section C, I examine sentences about 1st-order models, and argue that the speaker/writer/thinker of those sentences cannot be my proto-self because my proto-self does not contain a model that gives meaning to those sentences. It follows that the speaker/writer/thinker of those sentences must be a different kind of self, which I call my *1st-order self*. Since my 1st-order self can speak/write/think sentences about my 1st-order model, it functions as a narrator of my 1st-order model of my world and sends neuromotor signals to my body that result in vocalizations or characters on paper (or a computer screen). I argue that this narrator function is a kind of behavioral rule, and that many other behavioral rules (such as getting a drink when thirsty) can be performed by my 1st-order self. The enormous power of these *1st-order behavioral rules* is briefly explored.

Section D examines the observation that I can speak/write/think meaningful sentences about my 1st-order self and other objects in my world, which implies that I have a model that contains my 1st-order self and other objects. I call such a model a *2nd-order model* of my world. In addition to rocks and dogs and my 1st-order self, this 2nd-order model contains objects that stand for other 1st-order selves such as you. The self that contains my 2nd-order model cannot be my proto-self nor my 1st-order self because neither contain a 2nd-order model for which the sentences are about. Therefore, the self that contains my 2nd-order model must be a different kind of self, which I call my *2nd-order self*. My 2nd-order self can think about how my 1st-order self and your 1st-order self interact. In particular, my 2nd-order self can imagine how you might see your world. This ability opens up a new frontier for exchange of information and goods, forward-looking decision making, and evolution of social norms.

Section E explores whether I have higher-order selves. Section F addresses several implications of my models of my self, namely communicating with other selves, interpreting my conscious experiences, and making choices. Section G concludes.

B. My Self as My Body and 1st-Order Models of My World

I have many models of my world that I use for different situations. For example, when I am ice skating and thinking about the motion of my legs, ankles and feet, I have a model of my body that has those body parts. Neuroscience has discovered convincing evidence that my brain (as well as the brains of many animals) maps the sensory data it receives into a model of my body (e.g. Bermundez, et. al., 1998; Damasio, 1999). When I think "my ankles are shaking", the phrase "my ankles" implicitly implies that those ankles are part of my body, so the referent of "my" is my whole body.[15] In other words, in these instances I identify my self as my whole body. I will call this sense of my self my ***proto-self***.

When referring specifically to my proto-self, to be perfectly clear, I will adopt the convention of using the subscripted pronouns "I_0", "my_0" and "me_0". For example:

I_0 am cold.

[15] In the phrases "my hand", "my coat", etc. "my" implies possession or belonging-to. In contrast, the phrase "my song" does not imply that I own that song; it merely points to a specific song among the set of all songs. Similarly, in the phrases "my whole body" and "my self", "my" is merely a pointer in the sense of *that* body (self) among the set of bodies (selves); it does not imply the existence of some entity that possesses that body (self).

My$_0$ ankles are shaking.
Stopping my heart from beating will kill me$_0$.

The subscript '0' is intended to be silent not spoken. When I use these pronouns without a subscript, they are intended in their everyday non-technical sense that I assume we share.

Each of these statements is about my proto-self *as a model* of my body, and it is this model that gives meaning to these statements. Just as "earth" can denote a model of the structure of the earth, or it can denote an object in a model of the solar system, "proto-self" can also denote a model of my body or an object in model of my world that contains objects external to my body. For example, when I am ice skating, and say or think "I am gliding over the ice", I have a model with the ice and immediate surrounds (such as other skaters and obstacles) as well as an object (my proto-self) gliding across the ice. I will call a model containing my proto-self as an object and objects external to my body a ***1st-order model of my world***.

The following statements

I$_0$ am gliding over the ice.
My$_0$ finger is pointing at a spider.
The sun is shining on me$_0$.

are about my 1st-order model of my world which gives meaning to each statement. Again, the subscripts are intended to be silent not spoken.

A proto-self is meant to be a simple concept of self. In particular, while a proto-self can be an object in a 1st-order model of the world, **a proto-self does not contain a model of the external world**. It is easy to see that this assumption is necessary to avoid an infinite regress as follows. If a proto-self contains a model of the world (call it M), then M must contain a proto-self which contains M; i.e. M must contain M. However, a finite model cannot contain itself, so M must be a transfinite model. Clearly a proto-self that contains such an M would not be a simple (let alone realistic) concept of self.

A 1st-order model could, but need not, coincide with science, with parts of science or even be compatible with science. Indeed, I undoubtedly have many models (as does science) with various domains which form the basis for interpreting sentences. For instance, when hearing or reading statements about God, I can use a hypothetical model in which there is an object that is the referent of "God" that helps me interpret the statements, without committing to the pragmatic value of that model. On the other hand, evolution will have disfavored models that were seriously disadvantageous to my ancestors' survival. Therefore, I believe my various 1st-order models of my world are approximately consistent with the laws of Nature at the human scale as currently understood by science. For example, the dynamic of my proto-self in a 1st-order model satisfies the principle of "local causality" - i.e. the reaction of my proto-self at time t does not depend simultaneously on events spatially

separated from my proto-self. In Einstein's words: "no spooky action at a distance."

Figure 1 is a suggestive diagram of a 1st-order model of my world with my proto-self, a ball, a star, and a dog that is barking.

Figure 1. A 1st-Order Model

This 1st-order model gives meaning to the statements:

A dog is barking. (A)

The model comes first, then the sentence about the model. Therefore, **to understand a sentence it is necessary to have a model that gives meaning to the sentence**. Fortunately, the context usually gives sufficient information to construct a model that is a reasonable approximation of the speaker's (writer's) model. However, confusion between speaker and listener is always possible.

Note that this requirement resolves Searle's paradox of the machine that translates Chinese into English. Searle argues that the machine does not understand the Chinese, even though the output of the machine in English may be identical to the output a human translator would generate. In contrast, replacing the

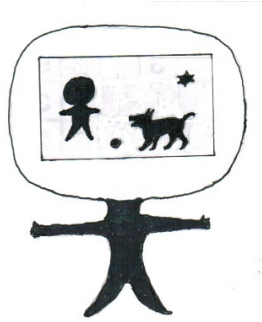
machine by a human that is fluent in Chinese and English, I argue that the human can construct a model that gives meaning to the Chinese input, and a corresponding model about which he/she can construct English sentences. The difference between the machine and the human translator is the existence of a model in the human that gives meaning to the sentences.

C. My 1st-Order Self and My 1st-Order Model.

Surely my 1st-order model of my world is part of my self. That is, my self contains a 1st-order model of my world which contains my proto-self. The self that contains this 1st-order model cannot be my proto-self, because my proto-self does not contain a model of my world. Therefore, the self that contains this 1st-order model of my world is *different in kind* from my proto-self. I emphasize "different in kind" to distinguish this difference from "difference in degree" which applies to the more or less detailed models of my body. I will call this different kind of self my ***1st-order self.***

Figure 2 is a suggestive diagram of a 1st-order self as a model.

Figure 2. A 1st-Order Self as a Model



In this diagram my 1st-order self *as a model* is depicted as a body-like figure with an enlarged head that contains my 1st-order model. The remaining (as yet undifferentiated) interior of the head contains, for example, neural mechanisms that can formulate a statement about my 1st-order model and send signals to other body parts that result in vocalizing or writing a statement about my 1st-order model, or other actions by my body. To be clear when referring specifically to my 1st-order self, I will adopt the convention of using the subscripted pronouns "I_1", "my_1" and "me_1"; again the subscripts are meant to be silent.

At this point some readers may expect me to revise the definition of a 1st-order model of my world by replacing the proto-self with a 1st-order self. However, to do so would lead to an infinite regress: my 1st-order model contains my 1st-order self which contains my 1st-order model which contains my 1st-order self and so on forever. To avoid this infinite regress, I stand by

the definition of a 1st-order model as containing my proto-self which does not contain a model of my world.

To make statements in which the referent of "I" is my proto-self (I_0), my 1st-order self (I_1) must contain or have access to my 1st-order model because such statements are about my 1st-order model. This requirement is suggestive of being conscious of my 1st-order model. This suggestion will be taken up in Subsection F.1.

In general, my 1st-order self has the ability to make statements about my 1st-order model. In this sense, my 1st-order self is a **narrator** of my 1st-order model. In addition to declarative statements such as sentence A, my 1st-order self can make statements about relationships in my 1st-order model, such as "the sun moves continuously across the sky from east to west", and "whatever goes up, (if unimpeded) must come down." These are basic physical relationships that evolution is likely to have hard-wired into our brains (1st-order models). Beyond these, given a memory of past states of my 1st-order model, it is possible for my 1st-order self to perceive temporal patterns and narrate them, such as "when a dog approaches a cat, the cat will run away", and "if it begins to rain and thunder, you_0 will take cover". Note that these statements are predictions about the future given the current state of my 1st-order model. This ability could differ across individuals (and species).

Moreover, given a prediction made in terms of a 1st-order model, the prediction is falsifiable by direct observation. Repeated falsifications could lead to modifications of my 1st-order model to improve its forecasting accuracy. Essentially, the scientific method could be hardwired into my 1st-order self.

In addition to a memory, my 1st-order self has a workspace in which it can simulate the future of my 1st-order model. Such a simulation would be an imagined world. Since an important component of such a simulation would be my_0 next action, by simulating the imagined future under alternative available actions, my 1st-order self can generate associated imagined future scenarios. Choosing which action will be addressed in Subsection F.3.

1. Who wrote sentence A ("a dog is barking")?

It is important to distinguish between (i) the string of characters comprising sentence A (or the sound waves if A is spoken), and (ii) the meaning of A in terms of a model of the world. We have become accustomed to mechanical devices such as personal computers and smartphones that display text messages and produce the sound waves but do not understand the meaning of the words. A computer program simply executes the rule: when asked question Q, reply with text or sound R(Q). In contrast, when I write (or say) sentence A, it has meaning in

reference to a model - in this case my 1st-order model that contains my proto-self, an object called a dog and other objects, and dynamic actions such as sound waves corresponding to barking.

Henceforth, I will adopt a **square-bracket convention** that the statement

(An object) wrote (or said) **[ ]**.

implies merely that *the characters (or sound waves) in square brackets [ ] were produced by the named object but not necessarily understood*.

Further, I will adopt a **curly-bracket convention** that the statement

(An object) wrote (or said) **{ }**.

implies that *the words in curly brackets { } have meaning in terms of the named object's model of its world*. In other words, putting square brackets around sentence A, namely [A], conveys that the sentence should be interpreted merely as a string of characters (or sound waves), whereas putting curly brackets around sentence A, namely {A}, conveys that the sentence has meaning in terms of the writer's (or speaker's) model of the world.

Accordingly, we can ask two kinds of questions about the source of sentence C.

Who wrote [A]? (Q1)

Who wrote {A}? (Q2)

Of course, in everyday language, we would ask simply "Who wrote A?", without the square or curly bracket conventions. However, the distinction is important, because "wrote" does not have the same meaning in Q1 and Q2. In Q1, "wrote" means merely that the characters in [ ] were "mechanically/unconsciously produced", while in Q2 "wrote" means that not only was the character string produced, it also had meaning in terms of the writer's model of the world. Henceforth, for extra clarity I will use the square-bracket and curly-bracket conventions.

For Q1, the answer must be an object in my 1st-order model, which could be a machine or my proto-self; e.g. "I_0 wrote [A]." For Q2, since a machine and a proto-self does not have a model of the external world, the answer to Q2 cannot be a machine or my proto-self. Hence, the answer to Q2 would be *nobody*. On the other hand, since I indeed wrote the character string [A] and I understood the meaning of [A] in terms of my 1st-order model, it follows that the answer to Q2 is my 1st-order self. In other words, we can transcribe A as:

I_1 wrote {a dog is barking}. (A')

It is a feature of the English language that a word can have very different meanings in different contexts. By transcribing

sentences using subscripts on pronouns and the square-bracket and curly-bracket conventions, as in "I_0 wrote [A]" and in "I_1 wrote {A}", hopefully the different meanings of "wrote" become clear.

Obviously the model that gives meaning to "I_0 wrote [A]" is my 1st-order model. But what is the model that gives meaning to "I_1 wrote {A}"? It cannot be my 1st-order model because my 1st-order model does not contain I_1. The answer is that my 1st-order self *as a model* gives meaning to "I_1 wrote {A}".

My 1st-order self as a model contains the physical boundary of my 1st-order self (i.e. my skin), my 1st-order narrator function, my 1st-order model of my world which contains my proto-self and other objects and relationships among these objects, a workspace for simulations, and other internal structures/functions, but it does not contain objects external to my 1st-order self.

My assertion that "I_1 wrote {A}" is true from my perspective; however, from your perspective it is not obviously true. Indeed, you may believe that I did not understand the meaning of the words, but instead simply scribed the character string [A]. In other words, you could believe that my proto-self or some other object in your 1st-order model wrote [A].

Instead of sentence A ("a dog is barking"), consider the seemingly equivalent sentence:

I hear a barking dog. (B)

What is the model that gives meaning to sentence B? It could be my 1st-order model, in which case the referent of "I" is my proto-self, and since my proto-self does not understand the meaning of words, to make this interpretation clear, we would transcribe B as

I_0 hear a [barking dog]. (B')

Note that the square bracket notation implicitly modifies the meaning of verb "hear". Specifically, it means merely that the incoming sound waves to my proto-self are associated in my 1st-order model with the label [a barking dog], in contrast to *perceiving* a barking dog. To be more specific about how my proto-self could correctly label the incoming sound waves [a dog is barking], suppose:

My proto-self has a file of sound waves indexed by a countable set of distinct labels. Given an incoming sound wave, my proto-self finds the best match in this file. The label associated with this best match is "a barking-dog". Then, "I_0 hear [a dog barking]" means merely that "the label associated with my proto-self's best match to the incoming sound wave is [a barking dog]. This meaning does not imply that my proto-self understands the words as my 1st-order self could. That is, B' is a statement about my 1st-order model, as is A, conveying the same information as A.

Alternatively, B could be a statement about my 1st-order self as a model, in which case the referent of "I" is my 1st-order self. My 1st-order self has a model of my world with a barking dog, which gives meaning to the words "a barking dog". Therefore, we would transcribe B as

I_1 hear {a barking dog}. (B")

in which "hear" has its ordinary connotation of perceiving the incoming sound waves as those of a barking dog.

In the previous analysis of "Who wrote A?" there were two meanings of the verb "wrote", and the square and curly bracket notation indicate which meanings. In the current analysis of "I hear a barking dog?" the bracket notation implicitly modifies the meaning of verb "hear".

2. More Sentences About My 1st-Order Model.

In this subsection I will analyze a collection of different kinds of sentences about my 1st-order model in order to illustrate how to transcribe them using subscripts on the pronouns and the square and curly bracket conventions. To start, consider the sentence:

I am anxious when riding in a car. (C)

Besides stating the act of riding in a car, this sentence asserts that the subject has the property of anxiety. At first one might

argue that since the proto-self is like a physical object, and physical objects do not have emotions, the subject cannot be my proto-self. However, just as electrons and neutrons are different kinds of physical object with unique properties, a proto-self is a kind of object and can have unique properties among which are "emotions". I put the word "emotion" in quotes because there are at least two meanings of the word: (i) a disposition or propensity and (ii) a feeling. By interpreting my proto-self as being anxious, I emphatically do not mean that my proto-self feels anxious or is conscious of being anxious. I mean merely that my proto-self has the disposition (equivalently propensity) to behave in particular ways. Analogously, a rock has the property of mass, which implies it will behave in particular ways, but the rock is not aware of its mass or equivalently its dispositions. Millions of years of evolutionary selection pressure have shaped these dispositions, but they are essentially fixed over the lifetime of a human. Accordingly, I interpret the "I" in sentence F as my proto-self (I_0). Hence, sentence C should be transcribed as

I_0 am anxious when riding in a car. (C')

Who wrote[16] {C'}? In accordance with my curly bracket convention, the writer must have understood the character string

[16] In order to avoid tiresome phrases, I will henceforth leave it to the reader to insert "said or thought". I also tire of writing "one of my 1st-order models", so henceforth whenever I write "my" 1st-order model, "one of my 1st-order models" should be understood.

[F]; otherwise, the answer is *nobody*. Since F is a statement about my 1st-order model, the answer cannot be my proto-self. Therefore, if there is a proper answer other than "nobody" it must be my 1st-order self (I_1):

I_1 wrote {I_0 am anxious when riding in a car.} (C")

You, the reader of this chapter, on the other hand may not reach this conclusion. To the question asked, you may believe the answer is *nobody*, because you believe I am a proto-self or some other object in your 1st-order model of your world that scribed the character string [I am anxious when riding in a car] [17]. Since I cannot provide you with proof that these statements came from my 1st-order self, you (the reader) are free to interpret any or all of my statements herein as merely character strings coming from an object in your model of the world.

Next, consider a sentence about a past event such as

I rode on a train last month. (D)

Let t_1 denote the time this sentence was spoken, and let $t_0 < t_1$ denote the time last month of the event 'riding on a train'. At time t_0, I could have said "I am riding on a train". Like sentence F, the referent of "I" in this imagined present-tense sentence would be my proto-self at t_0 and the implicit speaker would be

[17] Note that since a character string (or sound wave) in itself has no meaning, there is no need to retain any subscript on "I".

my 1st-order self (I_1) at t_0. To make this clearer, we can use the symbols $I_0(t_0)$ and $I_1(t_1)$ etc., to make the time explicit. Accordingly, $I_1(t_0)$ would have been the speaker of "$I_0(t_0)$ is riding on a train". At time t_1, sentence D still conveys information about my proto-self at t_0, so the speaker of D must be my 1st-order self at t_1. $I_1(t_0)$ and $I_1(t_1)$ are the same kind of self (i.e. 1st-order), just at different times; that is

$I_1(t_1)$ wrote {$I_0(t_0)$ rode on a train at time t_0}. (D')

Next, consider a statement of denial/confession:

I lied when I said {I am hungry}. (E)

Similar to the analysis of sentence D, we can interpret "hungry" as a disposition of my proto-self, so the third "I" in sentence H refers to my proto-self. Recalling that the curly bracket convention implies that the speaker of those words understands their meaning in terms of a model (namely, my 1st-order model), the second "I" refers to my 1st-order self; i.e. "I_1 said {I_0 am hungry}".

The referent of the first "I" in sentence E is not immediately obvious. It might seem that for the same reasons that the second "I" cannot be my proto-self, the first "I" cannot be my 1st-order self. On the other hand, the self who said "I am hungry" is also the self who lied. Indeed, "lied" is merely a clarification of the verb "said", so it must be that the first "I" in H is also my 1st-

order self as a model. Hence, sentence H should be transcribed as

I_1 lied when I_1 said {I_0 am hungry}. (E')

The model that gives meaning to H' is my 1st-order self as a model.

As another example, consider the sentence about a belief:

I believe I have the flu. (F)

The embedded sentence of J (I have the flu) can be interpreted as {I_0 have the flu}. Since my proto-self cannot make statements about itself, the referent of the first "I" in J must be my 1st-order self (I_1). That is,

I_1 believe {I_0 have the flu}. (F')

That I_1 *believe* the embedded sentence is an assertion about the confidence I_1 have in the veracity of the embedded sentence. It allows for the possibility that the embedded sentence might not be true (i.e. that I_0 do not have the flu). It also suggests that I do not have sufficient evidence that proves I have the flu. In other words, sentence J declares a relationship between my 1st-order self and a statement about my proto-self.

Next, consider the counterfactual statement.

I wish I had said [thank you]. (G)

This is a statement about my 1st-order model, albeit about what could have been. By my convention, the referent of the second "I" in G is my proto-self counterfactually producing the sound waves [thank you]. Implicit in G is the memory that I_0 did not say [thank you] and the imagined counterfactual that I_0 did say [thank you]. These two possible states of my 1st-order models are still elements of my 1st-order self as a model. Therefore, the referent of the first "I" in G is I_1. That is,

I_1 wish {I_0 had said [thank you]}. (G')

Obviously, such a 1st-order self is more sophisticated than the foregoing examples; however, this difference is a matter of degree/detail and not a difference in kind. Also note that G' is a *judgment* about a 1st-order model; in other words, my 1st-order self can make judgments about my 1st-order model.

Finally, consider the proposition:

If I am hot and thirsty, I like a cold drink. (H)

A proposition consists of a condition and an implication. Both the conditional part and the implication part contain word strings that could stand alone as declarative sentences. By the same logic as applied above, the "I" in both of these embedded sentences refers to my proto-self (I_0). Sentence L asserts a relationship between two states of I_0: (i) hot and thirsty, and (ii) like a cold drink. Since my 1st-order model contains my proto-self, other objects, their properties and the relationships between

them, it can contain relationships like H. Since my 1st-order self (I_1) can make statements about my 1st-order model of the world, I_1 can write and understand sentence H. Therefore, the writer of sentence H is my 1st-order self:

I_1 wrote {if I am hot and thirsty, I like a cold drink}. (H')

3. 1st-Order Behavioral Rules.

Implicit in the conclusion that my 1st-order self can state a proposition like H is the ability to generate the characters or utter the sound waves [if I am hot and thirsty, I like a cold drink]. In other words, my 1st-order self can send signals to my body parts (fingers, lungs, vocal cords and mouth) that result in the production of the characters or sound waves. Functionally, such a neural-muscular mechanism that produces sentences about my 1st-order model is a ***narrator*** for my 1st-order self.

Given the ability of my 1st-order self to produce characters and sound waves corresponding to sentences about my 1st-order model, it is a small step to assume that my 1st-order self can implement a ***1st-order behavioral rule*** such as

If I_0 am hot and thirsty, then get a cold drink. (J) [18]

[18] To avoid confusing "I" as the first person pronoun and (I) as a sentence, I have skipped the letter I to denote a sentence.

'Hot and thirsty' is a property of my proto-self, and properties of my proto-self can be detected by my 1st-order self. 'Get a cold drink' is an action between two objects (my proto-self and a cold drink) in my 1st-order model. All that is required is for my 1st-order self to detect that 'I_0 am hot and thirsty' (which I_1 obviously can since I_1 can make statement J) and to send signals to a neural-muscular mechanism that result in getting a cold drink. More generally my 1st-order self has the capability to implement billions or even trillions of 1st-order behavioral rules like J. Note that my 1st-order behavioral rules, including my 1st-order narrator, are elements of my 1st-order self as a model

For illustration purposes, I could also have the rule: "If I_0 am hot and thirsty, then get a hot coffee." This rule obviously conflicts with J, and has a much lower propensity to be executed than J. Accordingly, to each 1st-order behavioral rule with the same condition, there is a strength or propensity to be executed. The greater the propensity of a rule, the more likely it will be executed.

The following is a simple trading rule:

If you_0 give me_0 X, then I_0 will give you_0 Y.

X could be a commodity or a favor, and Y could be money, an IOU or a promise to return the favor.

The power that 1st-order behavioral rules give my 1st-order self cannot be overemphasized. While the dynamics implicit in

my 1st-order model can entail inertia, simple phobias and mechanical stimulus-response functions, a 1st-order behavioral rule could override these autonomous responses. Since 1st-order behavioral rules can take as input the whole state of my 1st-order model which has been generated by a continuous influx of sensory information, the rules can be astronomically more sophisticated/complex than local stimulus response functions. By local I mean the response is a function only of the currently arriving sensory information. For example,

- When approaching a blind street corner, prepare to avoid oncoming traffic.
- If there are two lines at the ice cream counter, get in the shorter line.
- When storm clouds approach, find shelter.

Moreover, equipped with a memory of the history of my 1st-order model, it is possible to have behavioral rules that depend on the history of the state of my 1st-order model. For example,

- You$_0$ helped me yesterday, so I$_0$ will help you$_0$ today.
- If you$_0$ have lied three or more times, don't believe anything you$_0$ say.
- If your$_0$ advice has been consistently right in the past, give your$_0$ future advice serious consideration.

I leave it to the reader to produce countless other examples.

D. 2nd-Order Models of My World and My 2nd-Order Self

In Section C, I argued that since I indeed wrote sentence A ("a dog is barking"), the referent of "I" is my 1st-order self. That is,

I_1 wrote {a dog is barking}. (A')

Recall that my 1st-order self *as a model* gives meaning to A'.

1. Who wrote sentence A'?

Consider the question in ordinary language: Who wrote sentence C'? This question could have two meanings:

Who wrote [A']?

Who wrote {A'}?

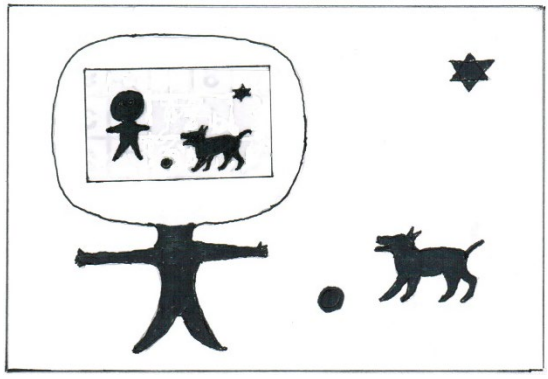
To be clear, the character string [A'] ≡ [I wrote 'a dog is barking']; where the subscript in A' has been dropped since it is not used in ordinary language, and the curly brackets around 'a dog is barking' have been dropped because the "I" in [A'] is only a character which obviously does <u>not</u> understand the meaning of the words 'a dog is barking' in terms of a model of the world. Consequently, the answer to Q3 could be a machine or my proto-self.

In contrast, the answer to Q4 cannot be a machine or my proto-self because those objects do not have models of the world

that would give meaning to A'. If the answer is not *nobody*, then it would have the form: "I wrote {A'}" = "I wrote {I_1 wrote {a dog is barking}}", where the referent of "I" is an object in a model that contains my 1st-order self as an object in its interior writing and understanding the words "a dog is barking".[19] Unfortunately, my 1st-order self is not an object in my 1st-order model.

Therefore, if the answer to Q4 is not *nobody*, I must have a higher level (different-in-kind) model of my world that contains my 1st-order self and objects external to my 1st-order self. I will call such a model a ***2nd-order model of my world***.

Figure 3. A 2nd-Order Model



[19] If instead of A', we considered "I_0 wrote [a dog is barking]", the answer to Q4 could be my 1st-order self.

Figure 3 is a suggestive diagram of a 2nd-order model. It contains a large body-like figure that stands for my 1st-order self as an object (see Figure 2), as well as other objects.

A 2nd-order model can also contain objects that stand for other 1st-order selves[20]; in such cases, for clarity I will use the generic "you$_1$", "he$_1$" or "she$_1$" to denote such a 1st-order self different from me$_1$. Inside the enlarged head is a diagram of my 1st-order model that contains a figure that stands for my proto-self as well as other objects.

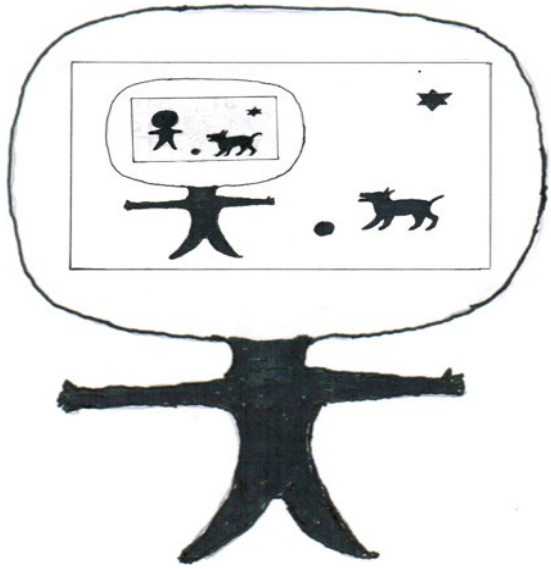
2. My 2nd-Order Self

Surely my 2nd-order model is part of my self. Moreover, this instance of my self must contain a 2nd-order model of my world which contains my 1st-order self. Since this self must contain my 1st-order self, the self that contains this 2nd-order model cannot be my 1st-order self. Therefore, the self that contains a 2nd-order model of my world is different in kind from my 1st-order self. I will call this new kind of self my ***2nd-order self.***

Figure 4 is a suggestive diagram of a 2nd-order self as a model.

[20] When I turn my attention to other mammals, I am willing to infer from neuroscience that many have a model of their body, and some may have 1st-order models of the world and 1st-order selves, but I am reluctant to assume that they have 2nd-order models of their world. Nonetheless, my 2nd-order model of my world can represent them as 1st-order selves, thereby allowing me to predict their behavior based on my hypotheses about their 1st-order behavioral rules.

Figure 4. A 2nd-Order Self as a Model



In this diagram my 2nd-order self is depicted as a body-like figure with an enlarged head that contains my 2nd-order model. The remaining (as yet undifferentiated) interior of the head contains, for example, neural mechanisms that can formulate a statement about my 2nd-order model and send signals to other body parts that result in vocalizing or writing a statement about my 2nd-order model, or implementing other actions.

To be clear when referring specifically to my 2nd-order self, I will adopt the convention of using the subscripted pronouns "I_2", "my_2" and "me_2". Again the subscripts are silent.

Thus, the answer to Q4, is either nobody or my 2nd-order self. Specifically, in the latter case, there are times $t_2 > t_1$, such that the answer can be transcribed as

$$I_2(t_2) \text{ wrote } \{I_1(t_1) \text{ wrote } \{\text{a dog is barking}\}\}.$$

Just as my 1st-order self as a model gives meaning to A', my 2nd-order self *as a model* gives meaning to A".

So, is the answer to Q4 nobody or is the answer my 2nd-order self? I have already asserted that when I wrote sentence A ("a dog is barking"), it was about a 1st-order model of my world containing a barking dog, so the writer could not have been my proto-self. Therefore, using subscripts and the curly-bracket convention: "$I_1(t_1)$ wrote {a dog is barking}" ≡ A'. But when I wrote A', did I understand the meaning of A'? Specifically, did I have a model of the world that contained my 1st-order self as an object and A' as the product of my 1st-order narrator? I have no doubt that I did since I can readily and easily write or say sentences about my 1st-order self (e.g. B', C', D', E', F', G', and H') and also sentences about other objects such as you_1, and this constitutes writing or saying sentences about a 2nd-order model of the world in contrast to a 1st-order model of the world. In conclusion, my answer to Q4 is my 2nd-order self.

You, the reader of this essay, may not agree with this conclusion. To the question asked, you may believe the answer is *nobody* because you believe I am a proto-self or some other

object in your 1st-order model of your world that wrote the character string [a'].

3. Further Benefits of a 2nd-order Self and a 2nd-order Model.

a. Models of Other Selves.

While the existence of my 2nd-order self is obvious (to me), surely there must be a greater purpose than answering questions like who wrote 'I wrote'. One of the most important advantages of a 2nd-order self and a 2nd-order model of my world is that the latter can also contain objects that stand for other 1st-order selves.

The difference between you_0 in my 1st-order model and you_1 in my 2nd-order model is that unlike you_0, you_1 contains my representation of $your_1$ 1st-order model. In other words, I have a Theory of Mind for you. Since I do not have direct access to $your_1$ 1st-order model, my_2 version is at best a guess based on the assumption that when you_1 appear in my 2nd-order model, $your_1$ 1st-order model is likely to be similar to my 1st-order model. Further, my version of you_1 will include presumptions about your 1st-order behavioral rules. Given these presumed behavioral rules and my version of your 1st-order model, my 2nd-order self can make predictions about $your_1$ behavior. For example, I_2 could believe that you_1 see me as an enemy instead of a friend. Obviously, this belief will influence how I_2 interact

with you$_1$. By observing past interactions, I$_2$ can update my belief about you$_1$ thereby modifying my 2nd-order model.

Moreover, since your$_1$ statements are about your$_1$ 1st-order model based on your sensory data (which is always different from mine), having a 2nd-order model with you$_1$ allows me to interpret your$_1$ statements and to supplement (as appropriate) the sensory data that goes into my models. When statements of many other 1st-order selves are slight variations of a consensus statement, and when such consensus statements in the past have been reliable, it will benefit me$_2$ to adjust my$_2$ model to be consistent. In this way, my 2nd-order self can benefit from a community of 1st-order selves far more than my 1st-order self can benefit from a community of proto-selves.

Another potential feature of my 2nd-order self is the ability to forecast the future of my 2nd-order model. My 2nd-order self could do this by identifying patterns from stored history of my 2nd-order model and extrapolating forward. In other words, beyond having a contemporaneous model of the world, my 2nd-order self has a dynamic 2nd-order model of the world. Since an important component of this dynamic model is the collection of my 1st-order behavioral rules, by simulating the future under the default and alternative 1st-order behavioral rules, my 2nd-order self can generate associated future scenarios. These future scenarios can influence the propensity to execute specific 1st-order rules.

It is important to understand that such evaluations of 1st-order behavioral rules cannot be performed by my 1st-order self since the later can only think (have an internal monologue) about my 1st-order model which does not contain my 1st-order behavioral rules. Therefore, deliberate modification of 1st-order behavioral rules based on expected performance is possible only with a 2nd-order self.

Consider again the 1st-order behavioral rule for trading:

If you$_0$ give me$_0$ X, then I$_0$ will give you$_0$ Y.

How would such a rule work in practice? In particular, after you$_0$ give me$_0$ X, what happens if I$_0$ do not give you$_0$ Y? The answer depends on the enforceability of my$_1$ promise, which depends on many details of my$_2$ model of the world dealing with the consequences of reneging on such a promise. To contemplate these issues – or to analyze any 1st-order behavioral rule - requires a 2nd-order model.

For example, I$_2$ could believe that "if you$_0$ give me$_0$ X, and I$_0$ do not give you$_0$ Y, then you$_0$ will hurt me$_0$." Equivalently, this behavior rule is included in my$_2$ model of you$_1$.[21] Clearly, if my$_2$ model of you$_1$ has such a rule, it will tend to deter me$_2$ from reneging. As a society, humans have developed extensive tort law to enforce mutually beneficial trade, and this could not have

[21] From your point of view this rule would be "If I$_0$ give you$_0$ X and you$_0$ do not give me Y, then I$_0$ will fight you$_0$.

happened without 2nd-order selves that can simulate 2nd-order models in which 1st-order selves have alternative 1st-order behavioral rules.

A 2nd-order self also empowers me to explain my behavior in terms of my 1st-order behavioral rules. For example, I_2 can write

> I_1 did Y because I_1 observed X and I_1 have the rule *if X then do Y*.

In contrast, I_1 could only write: "I_0 did Y because I_0 was in state S and I_0 have the disposition to do Y in state S."

Similarly, a 2nd-order self also empowers me to explain your behavior in terms of your 1st-order behavioral rules. For example, I_2 can write: "you_1 said 'thank you' because I_1 did you a favor and you_1 have the 1st-order behavioral rule *if someone does you a favor then say 'thank you'* ".

A 2nd-order self also empowers me to make judgments about 2nd-order models. As already noted, I_2 can compare two 1st-order behavioral rules and rank one better than the other. For example,

"If you_0 did me_0 a favor and now ask for a similar favor in return, I_0 will comply" *is a better rule than* "If you_0 did me a favor and now ask for a similar favor in return, I_0 will not comply."

I_2 prefer a 2nd-order model in which you_1 have the rule "if I_0 did you_0 a favor and now ask for a similar favor in return, you_0

will comply" to a 2nd-order model in which you_1 have the rule "if I_0 did you_0 a favor and now ask for a similar favor in return, you_0 will not comply" ceteris paribus.[22]

I cannot overstate the advantage of learning from others that is possible for a 2nd-order self. For example, I_2 could read (or hear) about a new (to me) 1st-order behavioral rule. Then I_2 could simulate my 2nd-order model under this new rule and compare the future scenario with that from my default rule. If the new rule is better than the default, then the propensity of the new rule will sharply increase; thus, better rules can be passed on by others (especially parents and mentors).

Closely related to learning new 1st-order behavioral rules is the possibility of improving my models of the world. I_1 could read (or hear) about alternative 1st-order models of the world: e.g. (i) that heavier objects fall faster than light objects, versus (ii) that all physical objects fall at the same rate on earth. By reading about Galileo's experiment, or performing a similar experiment and observing the result, I_2 can reject (i) in favor of (ii), and incorporate this new observation into my 1st-order model of the world.

[22] Both rules are stated from my_2 perspective. From $your_1$ perspective, $your_1$ rule would be stated "If you_0 did me_0 a favor and now ask for a similar favor in return, I_0 will (not) comply."

b. Second-order Behavioral Rules.

In addition to 2nd-order models of my world, my 2nd-order self can have *2nd-order behavioral rules* that are conditioned on the state of my 2nd-order model. This functionality gives me the ability to learn, communicate and interact in increasing complex ways.

For example, if my 1st-order self can execute a billion 1st-order behavioral rules, my 2nd-order self can execute on the order of a billion billion behavioral rules.

To appreciate the benefit of 2nd-order behavioral rules, consider the 1st-order behavioral rules for trading:

If you_0 give me_0 X, then I_0 will give you_0 Y.

How would such a rule work in practice? In particular, after you_0 give me_0 X, what happens if I_0 do not give you_0 Y? The answer depends on the enforceability of my_1 promise, which depends on many details of my_2 model of the world dealing with the consequences of reneging on such a promise. To contemplate these issues – or to analyze any 1st-order behavioral rule - requires a 2nd-order model.

For example, I_2 could believe that "if you_0 give me_0 X, and I_0 do not give you_0 Y, then you_0 will hurt me_0." Equivalently, this

behavior rule is included in my_2 model of you_1.[23] Clearly, if my_2 model of you_1 has such a rule, it will tend to deter me_2 from reneging. As a society, humans have developed extensive tort law to enforce mutually beneficial trade, and this could not have happened without 2nd-order selves that can simulate 2nd-order models in which 1st-order selves have alternative 1st-order behavioral rules.

Humans have developed informal and formal ways to summarize the history with other humans into an index of trustworthiness. For example, let Trust(you_1) denote my_2 index of trustworthiness of you_1, and assume I_2 update Trust(you_1) as follows: every time you_1 fulfill a promise, I_2 increase Trust(you_1) and vice versa. Note that the you_1 cannot be changed to you_0 because a "promise" implies that the meaning of the words is understood by the promiser and hence only a 1st-order self can make a promise. Let T* denote a positive threshold for trustworthiness. Then, the following pair of 2nd-order behavioral rules are implementable.

> If you_1 ask for X and Trust(you_1) ≥ T*, then I_1 will give you_1 X in exchange for $your_1$ promise to give me_0 Y by (date/time);

[23] From your point of view this rule would be "If I_0 give you_0 X and you_0 do not give me Y, then I_0 will fight you_0.

If you_1 ask for X and Trust(you_1) < T*, then I will refuse $your_1$ request.

Analogously, by examining my_2 memory of interactions with you_1 and him_1, I_2 could infer that you_1 are a better person (more trustworthy, kind, generous, etc.) than he_1 is.

E. A Higher-Order Selves?

The following sequence of ordinary language sentences raises the possibility of 3rd-order (or higher) selves.

a dog is barking. (A)

I wrote ‘a dog is barking’. (A’)

I wrote “I wrote ‘a dog is barking’ ”. (A”)

I wrote ‘“I wrote “I wrote ‘a dog is barking’ ” ”’. (A”’)

What are the referents of the I’s in sentences A’ to A”’? There are multiple possibilities.

(1) I could have used a recursive algorithm within my proto-self to generate these sentences, in which case all the characters after the first “wrote” in these sentences are simply characters, and the referent of the first “I” is my proto-self (I_0). In other words, A’ to A”’ would be transcribed as:

I_0 wrote [a dog is barking].

I_0 wrote [I wrote ‘a dog is barking’].

I_0 wrote [I wrote “I wrote ‘a dog is barking’ ”].

It is important to point out that the above three statements are the outputs of a *transcription algorithm*. Therefore, if you ask for whom are these transcriptions statements about a model of the world, the answer would be *nobody*.

(2) However, I have previously asserted that when I wrote A, it was a statement about my 1st-order model of the world. Further suppose statements A’ to A”’ are also statements about my 1st-order model. Hence, all characters after the second “wrote” in A” and A”’ are simply characters, the referent of the first “I” is my 1st-order self, and the referent of the second “I” is my proto-self. Then A’ to A”’ should be transcribed as:

I_1 wrote {a dog is barking}.

I_1 wrote {I_0 wrote [a dog is barking]}.

I_1 wrote {I_0 wrote [I wrote ‘a dog is barking’]}.

Unlike case (1), my 1st-order self as a model gives meaning to these transcriptions.

(3) However, in addition to (2), I have previously asserted that when I wrote A’, it was a statement about my 2nd-order model. Further suppose statements A” and A”’ are also

statements about my 2nd-order model. Hence, all characters after the third "wrote" in A" and A"' are simply characters, the referent of the first "I" is my 2nd-order self, the referent of the second "I" is my 1st-order self, and the referent of the third "I" is my proto-self. Then A" and A"' should be transcribed as:

I_2 wrote $\{I_1$ wrote $\{$a dog is barking$\}\}$.

I_2 wrote $\{I_1$ wrote $\{I_0$ wrote [a dog is barking]$\}\}$.

My 2nd-order self as a model gives meaning to these transcriptions.

(4) In addition to (2) and (3), if I were to assert that when I wrote A"', it was a statement about my 3rd-order model, then A"' should be transcribed as:

I_3 wrote $\{I_2$ wrote $\{I_1$ wrote $\{$a dog is barking$\}\}\}$.

However, I **cannot** honestly assert that I have a 3rd-order self that contains a 3rd-order model. While I can easily draw a modification of Figure 2 (with a 2nd-order self replacing the 1st-order self) thereby illustrating a 3rd-order model, such a 3rd-order model, in contrast to my 1st and 2nd-order models, is merely the output of a mechanical recursive algorithm.

If I do have a 3rd-order self, then I should be able to make meaningful statements about my 2nd-order self such as "I like you", transcribed as

$\{I_2$ like $you_2\}$ written by me_3. (K)

On the other hand, there are alternative transcriptions of "I like you", such as

$\{I_1$ like $you_1\}$ written by me_2. (K')

$\{I_0$ like $you_0\}$ written by me_1. (K")

[I like you] scribed by me_0. (K"')

To go from K"' up to K" entails that the source of 'I like you' understands the meaning of the statement in terms of the source's 1st-order model (e.g. $your_0$ physical features are pleasing to me_0), in which case the source is me_1. Similarly, to go from K" up to K' entails that the source of 'I like you' understands the meaning of the statement in terms of the source's 2nd-order model in which I and you are 1st-order selves (e.g. $your_1$ behavior is respectful of me_1) in which case the source is me_2. Both of these steps are self-evident to me.

However, the step from K' to K is not self-evident. In particular, I do not understand the meaning of the relationship "like" as applied to 2nd-order selves, which implies that I do not have a 3rd-order model with you_2 and me_2 and a relationship "like" applied to 2nd-order selves, so I am not a 3rd-order self.

As a byproduct of this analysis, we have uncovered an algorithm for interpreting any series of "I wrote 'I wrote …' " sentences without invoking higher than 2nd-order selves.

Specifically, given I have a 2nd-order self, transcribe the first "I" as my 2nd-order self (I_2) and enclose the remaining characters with curly brackets, transcribe the second "I" as my 1st-order self and enclose the remaining characters (except the terminal curly bracket) with curly brackets, transcribe the third "I" as my proto-self (I_0), and enclose the remaining characters (except the terminal curly brackets) with square brackets.

F. Further Implications of My Models of My Self.

1. Communication.

Communication between my self and other selves provides the opportunity to share information, assuming there is sufficient trust. It would not be an overstatement to say that communication is a major milestone in human evolution. Hence, it is necessary to understand how communication is possible within the models of self developed herein, especially since the other selves are of a different (lower-order) kind. I can be the recipient of potential information by hearing or reading sentences produced by other selves in my model of my world. Subsection (a) below addresses how I interpret those sentences. Subsection (b) addresses the related issue of why I might ask a question of some other self. Subsection (c) addresses the dual issue of why I might answer a question asked by some other self.

a. Interpreting Sentences I Read (or Hear).

Suppose sentence A ("a dog is barking") was not written by me, but instead was read by me.[24] How I interpret A depends on the source of A. At the very least, I need a model of my world and an object in that model capable of generating the character string [A]. Obviously my proto-self cannot interpret A because my proto-self does not have a model of my world. On the other hand, my 1st-order self could have an object in my 1st-order model capable of generating [A]. It could be a proto-self (you_0) or a non-human object such as Siri. However, in either case, even though I_1 can understand the meaning of A, I_1 cannot infer that the source of the [A] understood the meaning of the words which I denote as{A}. Nonetheless, I_1 may believe through experience that [A] contains useful information even though not understood by you_0. In other words, I_1 can rate an autonomous input-output function for reliability.

To infer that the source of sentence A understands the meaning of A, the source in my model would have to be a 1st-order self (say you_1), and hence I would have to be a 2nd-order self with a 2nd-order model that contains you_1 writing {A}. In this case, I_2 can interpret the sentence as {A}. However, if the source is you_1, since you_1 have a model in which there is no other self that could understand A, why would you_1 bother

[24] An analysis similar to what follows applies to the case in which I hear the words.

writing A? Perhaps you_1 believe through experience that I_0 will react (mechanically) in a manner that is beneficial to you_1. On the other hand, if you_1 believe I_0 will react to [A] in a manner that is beneficial to you_1, then you_1 have an incentive to write A even when it is not true. Therefore, I_2 have no guarantee that $your_1$ sentence A means {A}. Nonetheless, I_2 may believe through experience and verification that you_1 are trustworthy.

In summary, I_0 could not discern the source of [A]. I_1 could identify the source as you_0 or some other object in my_1 1^{st}-order model but the source would not understand the meaning of the words; nonetheless [A] could convey useful information to me_1. I_2 could identify the source as you_1 who understand the meaning of the words with the veracity and usefulness depending on my_2 past experience with you_1.

b. Asking and Answering Questions.

Suppose I hear something that sounds like a dog barking, but I am not sure it is really a dog. I might ask "Is that a dog barking?" Implicit in this question are possible models of my world: one with a dog barking, and other models without a dog barking but something else producing a sound similar to a dog barking. In other words, *this question is about models of my world*. More specifically, among my models that entail a sound similar to a dog barking, is any model more plausible than a model with a dog barking? The self asking this question cannot be my proto-self, because a proto-self does not have a model of

its world. On the other hand, the self asking could be my 1st-order self or my 2nd-order self.

To fully analyze the question "Is that a dog barking?" it is necessary to know to whom or what am I addressing this question?

First, perhaps the question is not being addressed to anyone. The question could be merely a way to acknowledge uncertainty or to mark the beginning of an investigation into the source of the barking sound in my 1st-order model. In either case, the question is *rhetorical* and addressed to no one and no object. Note that all of my acts of searching my models could be marked by vocalizing or thinking such a rhetorical question and seeking its answer.

Second, suppose the question is addressed to an object in my 1st-order model. Since the invention of the Internet and search engines such as Google, we have become accustomed to addressing questions to a machine which can be represented as an object in a 1st-order model. For questions about facts, search engines use brute force algorithms to find possible answers to our questions. In other words, the machine does not have a model of the world by which it can interpret our questions and deliver a sensible answer. These algorithms produce sufficiently relevant responses that make it worthwhile to engage them. In the future, artificial intelligence (AI) might develop models of the world that are capable of understanding our questions and

delivering sensible answers, similar to other human selves. When technology reaches this capability, we will need to modify our models of the world to include such AI objects. However, it would be unreasonable to assume that my 1st-order model could contain objects with models of the world when no other object in my 1st-order model contains models of the world. Therefore, it would be my 2nd-order model that would be modified to include such advanced AI objects. Another possible object in my 1st-order model is an object I identify as you_0, where you_0 are a proto-self with no model of the world. This case can be treated the same as addressing a question to a machine.

Third, suppose the asker is my 2nd-order self and I_2 am addressing this question to a human in my 2nd-order model. If I am sincerely addressing this question to a human whom I expect to understand the question, that human must have at least a 1st-order self (call it you_1). Since you_1 have a 1st-order model of $your_1$ world, the object in $your_1$ 1st-order model that stands for me is a proto-self, call it $your_1(me_0)$. Thus, $your_1$ answer to my_2 question would be addressed to $your_1(me_0)$ which is distinct from me_0 and obviously different-in-kind from me_2. Hence, $your_1$ reply to $your_1(me_0)$ may not be useful to me_2. Moreover, you_1 may have an incentive to reply with mis-information in order to induce $your_1(me_0)$ to take an action that is better for you_1 than the action you_1 believe $your_1(me_0)$ would take given a truthful reply. Further, you_1 believe I am a proto-type, namely $your_1(me_0)$ which by definition is incapable of understanding the reply, so

you$_1$ may not reply at all. On the other hand, if you$_1$ reply on many occasions, with experience I_2 may find that your$_1$ replies are typically useful, so it could be worthwhile to ask you$_1$ "Is that a dog barking?"

Fourth, suppose the question is addressed to my self. Which self?

My 1st-order self has a 1st-order model with a proto-self, so perhaps my 1st-order self could address the question to my proto-self. But my proto-self does not have a model of my world and thus could not understand or answer the question. Hence, this and (by the same reasoning any) question addressed to my proto-self is rhetorical.

My 2nd-order self has a 2nd-order model containing my 1st-order self, so perhaps my 2nd-order self could address the question to my 1st-order self. My 1st-order self could (i) understand the sound waves [Is that a dog barking?] coming from me$_0$ in my 1st-order model, (ii) confirm whether or not the barking sound is coming from a dog in my 1st-order model, and (iii) utter an appropriate answer. But I_2 can directly observe whether or not the barking sound is coming from a dog in my 2nd-order model, so it would be pointless for I_2 to ask me$_1$. Therefore, the simpler interpretation is that the question, and by the same argument, all questions that appear to be to one's self, are rhetorical.

Let's switch places and consider being asked "Is that a dog barking?" My 1st-order self could conjure models in which (i) the sound waves [Is that a dog barking?] emanate from an object (call it you_0), and (ii) there is or is not a dog barking in my 1st-order model. Therefore, I_1 could interpret the sound waves as a question about a 1st-order model. On the other hand, since you_0 have no model of the world, the question cannot be about $your_0$ model. Could the question be about my_1 model? Clearly, I_1 can verify whether or not my_1 1st-order model contains a dog barking. Thus, I_1 can understand the question and could provide an answer that reflects (is about) my 1st-order model. But why would I_1 bother to answer, since my_1 object for you is a proto-self (you_0) that cannot understand my answer? Perhaps I_1 believe there is the possibility that you_0 might have mechanical responses that could have favorable consequences for me_1. But then answering truthfully is not necessarily optimal. On the other hand, experience might allow me_1 to assess whether or not it is useful to answer you_0. Further, it is possible that AI could develop the ability of asking me questions that, if truthfully answered, lead to outcomes that are beneficial to me_1. Thus, 1st-order selves might answer questions posed by a machine (as well as a proto-self) if doing so is reliably beneficial to me_1. Obviously, experience could lead to the opposite conclusion, and therefore not answer or answer untruthfully.

If I am a 2nd-order self, then I_2 could have a 2nd-order model in which you_1 ask "Is that a dog barking?", and you_1 can

interpret my answer in terms of your$_1$ 1st-order model. In this case, I$_2$ might want to provide an answer (whether truthful or not) that elicits the action by you$_1$ that is best for me$_2$. On the other hand, if I$_2$ anticipate interacting with you$_1$ many times in the future, I$_2$ could forecast the negative effect of lying now on those future interacts and decide instead to answer truthfully. There is also the futuristic possibility that the question is being asked by an advanced AI machine that contains many models of its world and seeks to gather additional information from me. I$_2$ could have a 2nd-order model containing such an AI object (similar to having a model with you$_1$). Of course, I$_2$ may or may not deem it appropriate to answer truthfully depending on my$_2$ past experience with this AI object and on how I$_2$ perceive the effect of lying now on my future interactions with this AI object.

2. Conscious Experiences.

Consider the sentence: "I see a {red apple}". The curly brackets denote that I understand the meaning of a 'red apple' in terms of a model of my world. Therefore, this sentence has the same meaning as the following statement:

I$_1$ am having a conscious visual experience of a {red apple}. (L)

Since my proto-self does not have a model of my world, the "I" in L must be at least my 1st-order self containing my 1st-order model of my world in which there is an object called a 'red apple' in a position such that light from that object is incident on

my retina. Note that my interpretation L is very different from representationalism which asserts that 'to have a conscious visual experience of a red apple' implies I am conscious of a noumenal thing-in-itself. In contrast, all I am asserting is that there is an object that I call a red apple *in my 1st-order model* of my world. I am not making a claim about noumenal reality. To be clear, it would be better to restate sentence L as "I am having a *red-apple conscious experience*" which designates the kind of conscious experience without suggesting the existence of a noumenal red apple.

As I have done with similar sentences in this Chapter, I can ask who wrote sentence L. It clearly could not be my proto-self nor my 1st-order self since neither have a model that contains my 1st-order self as an object. Therefore, the answer must be my 2nd-order self: i.e.

I_2 wrote {I_1 had a conscious visual experience of a {red apple}}. (M)

Moreover, my writing M was preceded by a conscious thought experience, namely

I_2 had the conscious thought experience of a {red apple}}. (N)

In other words, my 1st-order self had the conscious sensual experience of seeing a red apple, while my 2nd-order self had the conscious thought experience of me_1 having a conscious visual

experience of seeing a red apple. My 2nd-order self as a model gives meaning to sentence N.

Some readers may notice a similarity between my analysis of the conscious experiences of my 1st-order and 2nd-order selves, and *higher order theories of consciousness*. E.g. see Carruthers (2000), and Gennaro (2004). However, unlike my analysis, those theories are built on a foundation of representationalism and reductive physicalism, and they do not use the concept of models of the world, nor the concept of different kinds of selves.

3. Making Choices.

When I face a situation in which there are two or more feasible actions available to me and only one action can be taken, how do I choose which action to take? For example, I enter an ice cream parlor, and need to choose from a dozen flavors offered. Or I need to choose a health plan from the many options. The process of making a choice can be simple or complex. For the ice cream example, the process may depend only on the immediate consequences and could be essentially automatic; i.e. my proto-self could have an inherent disposition for flavors. In contrast, for the health plan example, the process is likely to depend on the anticipated future consequences which I ascertain by simulating my model of the world into the future under the alternative health plans.

By some estimates (e.g. Szegedy-Maszak, 2005), 95% or more of my choices are made automatically by genetically determined instincts or by my autonomic nervous system. In my 1st-order model these choices are attributed to my proto-self. I could say "I choose to breathe faster when running", but it would be strange because in ordinary English "choose" does not apply to autonomic actions. When referring to an autonomic action taken by my proto-self, I simply say "I_0 breathe faster when running."

When facing choices that are not autonomic and have future consequences, the choice process can involve simulations of my 1st-order model into the future. For each available action, my 1st-order self can run a simulation in a workspace of my brain which gives the action a *utility value* as a function of the output of the simulation; then after the final simulation, I take the action with the highest utility value. For brevity, call this action *best*. Then I could say "I_1 choose the best action." Note that here "choose" does not imply a role for *free-will* because this "1st-order choice process" is the outcome of a biomechanical algorithm in my brain.

If I_2 believe I am interacting with a 1st-order self such as you_1, then the choice process could involve simulations of my 2nd-order model. For example, suppose I_1 have two available actions and you_1 have three available actions. Then, for each of my two available actions, I_2 would run three simulations of $my_2(your_1)$

1^{st}-order model, one for each of $my_2(your_1)$ available actions under the assumption that you_1 believe I am a proto-self that follows a stimulus-response function, and I_2 could store $your_1$ best action from $my_2(your_1)$ perspective, so after the last run, I_2 will have a prediction of how you_1 will respond to each of my available actions. After all these six simulations, I_2 will have identified my_2 best action and take that action. In other words, "I_2 choose the best action." As with the previously described "1^{st}-order choice process", here "choose" does not imply classical free-will because this "2^{nd}-order choice process" could be the outcome of an algorithm that is activated in my brain. On the other hand, to the extent than the action I_2 take is not completely determined by variables external to my 2^{nd}-order self – that is, my 2^{nd}-order process has an essential causal role – I could say that my 2^{nd}-order self is free from total determination by external causes.

Clearly, this "2^{nd}-order choice process", in this example, would take much longer and use much more brain resources than the "1^{st}-order choice process". Consequently, the 2^{nd}-order choice process will be reserved for choice problems that are perceived to have potentially significant consequences.

Consider also choices that have ethical content, such as

I care about my wife, my friends and my community.

What do I mean by "care" if it does not entail freely choosing to care, and what is the referent of "I" in this statement? The following are potential answers.

"I_0 care about my wife" means that my proto-self has a positive disposition towards my wife as an object.

"I_1 care about my $wife_0$" means that I have 1^{st}-order behavioral rules that act in caring ways towards my wife as a proto-self; for example, I_1 act in ways that increase her_0 happiness.

"I_2 care about my $wife_1$" means that I have 2^{nd}-order behavioral rules that act in caring ways towards my wife as a 1^{st}-order self; for example, I_2 encourage her_1 to adopt 1^{st}-order behavioral rules that improve her_0 health.

But why do I care?

"I_0 care about my wife" because evolution has resulted in my having dispositions towards females with sensual features like my wife.

"I_1 care about my $wife_0$" because I_1 have adopted 1^{st}-order behavioral rules that are positively correlated with successful marriages.

"I_2 care about my $wife_1$" because it benefits me_2 when my $wife_1$ adopts 1^{st}-order behavioral rules that improve her_0 health and happiness.

In other words, free-will is not necessary for caring about my wife because evolution has favored caring dispositions and caring-about-my-wife behavior.

G. Conclusions.

I began this Chapter with the question "What is my self?" I have maintained that *a statement by me has meaning only in reference to a model of my world*. To explore the implications of this premise for the concept "my self", I examined many sentences entailing the personal pronoun "I". This examination uncovered three kinds of models and three kinds of objects in those models as the referent of "I".

i) *My proto-self as a model* of my body but without a model of the external world; and *my proto-self as an object* in my 1st-order model of my world.

ii) *My 1st-order self as a model* which contains a narrator function, other 1st-order behavioral rules, and a 1st-order model of my world containing my proto-self as an object and objects external to my proto-self; and *my 1st-order self as an object* in my 2nd-order model of my world.

iii) *My 2nd-order self as a model* which contains a narrator function, other 2nd-order behavioral rules, and a 2nd-order model of my world containing my 1st-order self as an object and objects external to my 1st-order self.

I could define - but could not assert that I have – a 3rd-order self with a 3rd-order model of my world that contains my 2nd-order self as an object.

When I turn my attention to other mammals, I am willing to infer from neuroscience that many have a model of their body, and some may have 1st-order models of the world and 1st-order selves, but I am reluctant to assume that they have 2nd-order models of their world and 2nd-order selves. Nonetheless, my 2nd-order model of my world can represent them as 1st-order selves, thereby allowing me to predict their behavior based on my hypotheses about their 1st-order behavioral rules.

The idea that *my self is an object in a model of my world* resolves the mind-body problem. Specifically, it avoids the notion that my self is a Cartesian mind different in kind from physical things. While some of the objects in my model of my world are related to each other in so-called "physical" ways (e.g. Newton's Laws of Motion), other objects are related to each other in non-physical ways such as definition, logic and mathematics.[25] Moreover, the model itself and all objects in it are abstract, independent of whether instantiated on paper, in digital bits, or in neural patterns. Nevertheless, my models

[25] I strongly oppose calling this view "property dualism" because the objects in a model do not have *inherent* properties; instead their behavior is a result of the relationships between the different kinds of objects.

would not have pragmatic value if there was no correlation between them and noumenal reality.[26]

Communication between my self and other selves is still possible, provided there is sufficient trust established by past experiences. Human progress that comes about via sharing of information would not be possible without sufficient trust. Unfortunately, history has also shown that trust can be lost as well as built.

Further, my models of my self provide a foundation and answer to:

Who is conscious of a red apple? [my 1st-order self]

Who is conscious of being conscious of a red apple?

[my 2nd-order self]

In other words, my 1st-order self, by virtue of containing a 1st-order model of my world, is conscious of things in that model; and my 2nd-order self, by virtue of containing a 2nd-order model of my world, is conscious of things in that model. Hence, in my definition of a conscious experience as the *perception* of a model

[26] Note that since noumenal reality is unknowable, the correlation cannot be quantified; we can only say that it is greater than zero for sufficiently many of my models to ensure my survival for a finite amount of time.

of my world, the "perceiver" is my 1st-order (or 2nd-order) self, as indicated above.

Finally, because the actions I choose (via my behavioral rules) depend on my models of my world, my survival depends on the reliability of those models. Therefore, to the extent that I use the scientific method to improve the reliability of my models, my chance of survival will increase. To the extent that I evaluate behavioral rules based on past consequences, my chance of survival will increase.

IV. Scientific Phenomenism and Quantum Mechanics

This chapter will argue that Scientific Phenomenism provides a consistent interpretation of Quantum Mechanics (QM) as a model of phenomenal reality. In particular, phenomenism resolves the controversial "measurement problem" in QM.

QM provides a method for predicting measurements at the subatomic scale. It entails a complex-valued *wavefunction* that obeys *Schrodinger dynamics*, and mathematical *operators* on that wavefunction that correspond to acts of "perfect measurement". An act of perfect measurement is an interaction/event that transfers information to a measurement device. The magnitude-squared of the wavefunction is interpreted as the probability of attaining a specific measurement outcome, or as the large-sample limit of the frequency of specific measurement outcomes.

QM predictions based on this interpretation have never been falsified by experiments. Nonetheless, QM suffers from two so-called measurement problems: (1) the collapse of the wavefunction upon a measurement, and (2) the role of consciousness in measurement.

A. The collapse of the wavefunction.

To illustrate the first problem, consider a cathode ray tube (CRT), such as in an oscilloscope or an old black and while television. At the narrow end of the CRT is a cavity with a narrow opening in the direction of the wide end (the screen), and in this cavity is a cathode that is heated to a temperature at which electrons fly away from the cathode through the opening in the direction of the positively charged screen that is coated with a phosphorescent substance. When an electron hits the screen a tiny flash of light is emitted at the location of the hit. Assume the cavity opening is so small that the rate of electron emissions from the cavity is low enough that different electrons produce separate observable flashes on the screen. A flash of light on the screen of the CRT is a measurement outcome interpreted as the location of the electron at the time of the flash. This description is consistent with classical physics.

The QM description is quite different. The QM model posits a complex-valued wavefunction that obeys Schrodinger dynamics. The electron moving from the cathode towards the CRT screen is replaced by the continuous space-time dynamics of the

wavefunction.[27] A tiny fraction of a second after the electron leaves the cathode, this wavefunction becomes concentrated at the CRT screen but spread diffusely over the screen. At the moment of the flash of light on the screen, information is transferred to the screen and to any observer that is present. At this moment, the wavefunction changes discontinuously to one in which it is no longer diffusely spread over the screen, but instead is concentrated at the location of the flash. In this sense, it is often said that the wavefunction "collapses" at the moment of measurement. This discontinuous change is a violation of Schrodinger dynamics.[28] Furthermore, if the wavefunction is a physical entity with energy distributed throughout the wave function, then the collapse represents an instantaneous transfer of energy across space at superluminal speeds, in violation of Special Relativity.

Instead, one can interpret the wavefunction as a *nonphysical* entity in an abstract model that can predict measurement outcomes. Moreover, the wavefunction (an element of abstract Hilbert space) is unobservable; only acts of measurement are observable.[29] The

[27] More precisely, the projection of that wavefunction onto the subspace for that electron.

[28] Of course, actual human measurements are imperfect and so it is conceivable that Schrodinger dynamics could be modified to incorporate imperfect measurement in a way that preserves continuity of the wavefunction. However, QM currently has no extended theory that does this.

[29] This perspective in not unique to QM. Economic theory posits a utility function that represents an individual's preferences over possible consumption bundles. This utility function is unobservable; only consumption by the individual is observable.

wavefunction obeys Schrodinger dynamics between measurements but not at moments of perfect measurement. As a nonphysical unobservable entity in a model of measurement outcomes, the so-called collapse of the wavefunction is not a physical event. It is an *updating of the model* given the information provided by the measurement outcome (e.g. the flash of light on the CRT). Hence, this discontinuous change is no more troublesome than the ordinary updating of a probability measure given new information via Bayes Rule.

B. The role of consciousness.

The second issue of the role of consciousness can be addressed in two ways. First, suppose the wavefunction is a physical entity, and that the conscious observation of seeing the flash of light on the CRT screen causes the collapse. To illustrate the absurdity of this supposition, assume a visual recording of the screen was made during the experiment, but not observed by any human until one year after the experiment. To suggest that this delayed conscious observation caused the earlier physical collapse of the wavefunction entails the absurdity of reversing the arrow of time.

Second, suppose the wave function is a nonphysical entity in an abstract model of measurement outcomes. As discussed in Chapter III, such a model is one of an individual's background models which are accessed via conscious thought experiences. From this viewpoint, the conscious sensual experience of seeing a flash of light on the CRT can cause an updating of the model consistent

with QM. The point is that this conscious experience does not cause a collapse of a physical wavefunction, only an updating of one's abstract model. As a consequence, different individuals (even different quantum physicists) can have different current models which entail different subjective beliefs about the wavefunction based on their unique history of conscious experiences. This conclusion is hard for many physicists to accept, because it implies that beliefs about wavefunctions are personal/subjective.

To illustrate the necessity of this conclusion, suppose Alice did observe the flash when it happened, but Bob only observes the recording one year later. Alice will immediately update her model, but Bob cannot update his model until one year later. Therefore, Alice and Bob will have different beliefs about the wavefunction (i.e. different models) in the interim. Further, these different beliefs can have real effects. For example, suppose Bob does not know that Alice has observed the outcome when it happened and suppose the flash appeared in the upper right quadrant of the screen. Then, Alice could offer Bob a bet with even odds that the flash will appear in the upper right quadrant of the screen, and Bob, thinking it is three times as likely to appear in the other three quadrants, would readily accept that bet (and lose).[30]

[30] Obviously, Alice would be guilty of deceit, which is why inside trading on the stock market is illegal.

Note that after one year when Bob observes the recording and updates his model back to the time of the experiment, his model will then agree with Alice's model, provided their prior models were the same. In other words, the sharing of verifiable information, as is practiced in science, will reduce the differences in subjective beliefs. Hence, the subjective nature of phenomenism does not preclude the emergence of very similar models with similar predictive accuracies, and the practice of the scientific method makes such emergence more likely than any other method.

In conclusion, the two measurement problems of QM vanish in the framework of scientific phenomenism. QM is an abstract model, and as such the so-called collapse of the wavefunction is not a physical event, but instead the "collapse" is an updating of a mental model of phenomenal reality given new information (conscious experiences). Since conscious experiences are subjective, beliefs about the QM model as currently held by an individual human are subjective. Further, since noumenal reality is unknowable, questions about how well any model of phenomenal reality corresponds to noumenal reality are categorically unanswerable and hence a waste of time. We can only compare the accuracy/reliability of models of phenomenal reality.

By recognizing that noumenal reality is unknowable, phenomenism dissolves the dualistic problem of how the "physical" (in the sense of noumenal things-in-themselves) and the

"mental" (i.e. phenomenal) interact. Further, Scientific phenomenism invokes the scientific method to compare the reliability of the predictions of one's models and to update one's beliefs about those models accordingly.

V. Beyond the Mind-Body Problem

What are the practical implications of Scientific Phenomenism and models of my world and my self? First and foremost, since my survival depends on the actions I take, and those actions are the output of behavioral rules which depend on my models, using the most reliable (i.e. statistically accurate) models is critical for my survival. Since using scientific models to make predictions can have life and death consequences, improving the predictive performance of these models is necessary and sufficient for the maximum chance of survival. This is the pragmatic value of science (in contrast to the vain pursuit of discovering noumenal reality). Science is a collection of *reliable models*. Thus, science progresses by discovering more reliable models, not by making claims about noumenal reality.

Arguments about which model is "right" are wasteful misguided hubris. For example, for a century, enormous intellectual resources have been wasted on the debates about the

meaning of quantum mechanics – particularly the "measurement problem". Chapter IV showed how scientific phenomenism dissolves these issues by recognizing that quantum mechanics (like all other science models) is a *model*: a collection of abstract objects, relationships between those objects, and a dynamic of change, not a representation of unknowable noumenal reality.

Because a society's survival depends on the social policies it adopts, society should evaluate social policies using the most reliable models. Therefore, the enormous energy wasted on the mind-body problem (in particular the hard-problem of consciousness) could and should be directed towards building more reliable models of the world.

The most pervasive and consequential social policy of modern societies is its legal system. Ideally, the legal system should discourage behavior that society deems as unacceptable (such as decreasing the likelihood of society's survival). A prevailing method for achieving this goal is punishment. However, there is little scientific evidence that the penal code succeeds in discouraging unacceptable behavior (even during the period of incarceration). Efforts at rehabilitation are minimal. Revenge often seems more a motive for punishment than reparation and rehabilitation. Clearly, we need more reliable and statistically accurate models for discouraging unacceptable behavior and encouraging acceptable behavior.

Another example of an ineffective social policy is the drug war. Data reject the model in which (a) drug users freely choose to use drugs, and (b) prison changes their self-destructive behavior. The alternative model in which drug users are addicted to the drugs and society provides effective treatment is promising but needs to be implemented on a much wider scale.

If my survival depends on having reliable and statistically accurate models, isn't selfish greed the optimal behavior? Indeed, there are hypothetical models of the world in which selfish greed is optimal, but those models fail to take account of the interdependencies between humans and the environment. The canonical example is the *tragedy of the commons* in which selfish greed leads to the degradation of common resources and consequently makes everyone worse off. Interdependency implies that the optimal societal path entails some cooperation. Hence, societies that use models that incorporate pertinent interdependencies and adopt societal rules that induce appropriate cooperative behavior will be far more successful than those that don't.

Some critics will claim that without a necessary role for free-will (which my models of self lack), no one can be held responsible for unacceptable behavior. We see this legal defense offered more and more as science identifies neurological causes for bad behavior. Quite to the contrary, when unacceptable behavior can be attributed to mostly internal forces, those internal forces

(my model and/or my behavioral rules) **are** responsible (i.e. the causal reason). The task for society is to modify or counter those internal forces. Punishment has been the traditional response, but better methods for discouraging bad behavior need to be added to the toolbox. In contrast, when unacceptable behavior can be attributed to mostly external forces (such as poverty or child abuse), the task for society is to mitigate those external forces or provide protection against them, and to help offset the damages to the victims. Caruso (2019) and others advocate the "public health" model.

Critics may also argue that without the sense of free-will, individuals will cease "Caring" (caring about others, behaving better, and engaging in scientific research). Hence, they argue that philosophies denying free-will are dangerous to the survival of humans. Quite to the contrary, as I argued in Chapter IV.F.3, free-will is not necessary for Caring. Evolution will favor Caring provided (i) interdependencies are significant and represented in human models of the world, and (ii) human evaluation of alternative behavioral rules favors survival. The first condition is ensured by the feedback mechanism between predicted (simulated) outcomes and realized outcomes that drives improvements in the reliability and statistical accuracy of my models of the world. The second condition is self-evident. Both the modelling and evaluation processes are hard-wired into most human brains, so free-will is not necessary for Caring behavior.

The science fiction literature has explored imagined scenarios of computers taking over the world, but now advances in Artificial Intelligence are making such scenarios look much less fantastical. With the recent advances in AI (e.g. ChatGPT and GPT-4), there has been much discussion of the potential and dangers of advanced AI (e.g. Hunt, 2023). One of the issues is whether future AI systems will be considered conscious (Huckins, 2023) and therefore entitled to rights originally intended only for humans (such as privacy, due process, free speech, etc.) especially if future androids have bodies externally indistinguishable from humans.

However, since conscious experiences are private 1st-person experiences, there is no objective scientific way to verify whether or not such an android is really conscious. Therefore, *being conscious* cannot be a verifiable requirement for having civil rights. The best we can do is to specify a list of observable characteristics and behaviors, such as being awake, being responsive to external stimuli, answering and asking sensible questions, having certain types of brain waves, etc., as necessary and/or sufficient to be entitled to specific civil rights. For example, our legal system provides a list of sufficient behaviors (such as murder) for humans to be denied certain rights.

Currently, society's default necessary condition to be entitled to human rights is *being human* (i.e. having human DNA). Hence advanced androids would not qualify for human rights. However, many fans of the TV series Star Trek would disagree, arguing that

the intelligence and human-like behavior of the android called "Data" should entitle him to at least some human rights, such as freedom from arbitrary termination, cruelty, assault, etc.

On the other hand, these fans would not want Data punished for bad behavior in the way humans are punished in our legal system. For example, suppose Data commits a heinous crime, and is found guilty by a jury and sentenced to prison for life (as would a human who committed the same crime). But Data is an android, a machine that can be diagnosed and possibly repaired as needed. What a waste of resources to confine Data to prison. Obviously, the rational intervention would be to diagnose Data and then (i) repair Data if feasible, or (ii) terminate Data if no repair is possible and if society would be at unacceptable risk without termination. If we implement the diagnosis-and-repair paradigm for androids, we would lessen the chances of an evil android taking over the world because such subversive behavior could be diagnosed and repaired or eliminated. A similar argument could be applied to non-android AI (such as ChatGPT).

The point is that how we treat advanced AI is a cost-benefit problem, not an issue of consciousness. Further, since the justification for punishment of unacceptable behavior assumes the behavior was a conscious choice by the perpetrator and that punishment will lessen the chance of further incidents of such behavior by the perpetrator, and since there is no scientific (i.e. 3rd-person) way to verify that the perpetrator made a conscious choice,

punishment (rather than diagnose-and-repair) cannot be scientifically justified – not for androids nor for humans.

References

Berkeley, G., A Treatise Concerning the *Principles of Human Knowledge,* 1710.

Barrett, L. F. "How the Mind is Made", MIT Technology Review, 124 (5), 8-11, 2021.

Bermudez, J., Eilan, N., and Marcel, A., The Body and the Self, Bradford Books, 1998.

Byrne, A., "What phenomenal consciousness is like", in R. Gennaro (ed.), *Higher-Order Theories of Consciousness,* John Benjamins, 2004.

Cargile, J., "The First Person," Symposion, **6**, 23-28, 2019.

Carnap, R, "Psychology in Physical Language" (1932), in A.J. Ayer (ed.), Logical Positivism, New York: The Free Press, 1959, pp. 165–198.

Carruthers, P., Phenomenal Consciousness, Cambridge University Press, 2000.

Chalmers, D., "Facing up to the Problem of Consciousness," J. of Conscious Studies, **2**, 200-219, 1995.

Damasio, A., The Feeling of What Happens, Chp. 5, Harcourt, Brace & Co.,1999.

Descartes, R., Meditations on First Philosophy, Book 6, 1641.

Gallagher, S., and Shear, J., Models of Self, Imprint Academic, (2000).

Gennaro, R. (ed.), Higher-Order-Theories of Consciousness, John Benjamins Press, 2004.

Goldman, B. "Sense of Self: The Brain Structure That Holds Key to 'I'," Neuroscience News, June 22, 2023.

Huckens, G., "Machines Like Us", MIT Technology Review, **126**, 30-37. Nov/Dec 2023.

Hume, D., Enquiries Concerning Human Understanding, 1748.

Hunt, T., "Here's Why AI May Be Extremely Dangerous--Whether It's Conscious or Not", Scientific American, May 25, 2023.

Kant, I., Critique of Pure Reason, 1781.

Kuhn, T., The Structure of Scientific Revolutions, 1962.

Mach, E., The Science of Mechanics, 1883; translated by McCormack, Chicago: The Open Court Publishing Co., 1919.

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R., and Friston, K., "Bayesian Inferences about the Self: A Review," Consciousness and Cognition, **25**, 67-76, 2014.

Metzinger, T., "Self Models", Scholarpedia, **2**(10): 4174, 2007.

Mill, J. S., A System of Logic, Book V, Chapter V, 1843.

Neurath, O. (1931), “Physicalism: The Philosophy of the Vienna Circle”, in R. Cohen, and M. Neurath (eds.), Philosophical Papers 1913–1946, Dordrecht: D. Reidel Publishing Company, 1983.

Parvizi, J., et.al. (2023), “Causal evidence for the processing of bodily self in the anterior precuneus,” Neuron, June 8, 2023.

Plato, The Republic, Book VII, 375 BC.

Searle, J., “Minds, Brains and Programs”, Behavioral and Brain Sciences, **3**, 417–57, 1980.

Sellars, W. F. Science, Perception and Reality, International Library of Philosophy and Scientific Method, London,1963.

Turing, A., "Computing Machinery and Intelligence", Mind, **49**, 433–460, 1950.

Ulanovsky, N., “Neuroscience: How Is Three-Dimensional Space Encoded in the Brain?” Current Biology, **21**(21), 886-888, 2011.

Zahavi, D., Subjectivity and Selfhood: Investigating the First-Person Perspective, MIT Press, 2008.

ABOUT THE AUTHOR

Dale O. Stahl received his B.S. and M.S. in engineering from the Massachusetts Institute of Technology and his Ph.D. in economics from the University of California at Berkeley. He is an Emeritus Professor at the University of Texas in Austin. His academic publications are in economic theory and game theory. He is perhaps best known for his "Level-n Theory of Bounded Rationality" from which Chapter IV herein has evolved.

You are cordially invited to send comments to

sciphenom@gmail.com

or to join the discussion forum at

https://sciphenom.com.