# Agents with Free Will: A Theory Grounded in Quantum Physics

**Kathryn Blackmond Laskey**
Department of Systems Engineering
and Operations Research
George Mason University
Fairfax, VA, USA
klaskey@gmu.edu

***Abstract***—Does free will exist? It feels that way. We experience choosing freely among different possible actions, and these choices seem to have effects in the world. Yet the mainstream view among scientists is that our choices are entirely a function of neurobiological processes unfolding according to the laws of physics. Our intentions, the argument goes, play no causal role in our actions except as a high-level description of complex, underlying, physically determined processes in our brains and bodies. If this mainstream view were overturned, the implications for human society and for artificial intelligence would be profound. This paper explores a scientifically well-founded theory in which intentionality plays a fundamental causal role in our behavior. We begin by defining a set of properties that formalize the concept of genuine free will. We then present a theory of agency that satisfies these properties and is fully consistent with the laws and precepts of quantum physics. Next, a roadmap is given for evaluating the theory. Finally, implications for science, engineering, and philosophy are discussed.

## INTRODUCTION

Imagine that you are feeling thirsty on a hot summer day. You see an ice-cold, sugary drink sitting enticingly on a nearby table. You can almost taste it. You wage an internal battle between your thirst and your recent resolution to cut down on sugar intake. You think about reaching over for the drink, but hesitate. Which do you choose? The delicious, thirst-quenching drink or your health? As you struggle with the decision depicted in Figure 1, it certainly feels as if both choices are physically possible. After you have chosen and acted on your choice, it feels as if it would have been possible to have chosen otherwise. It feels as if *you* are the one who made the choice and caused the outcome. Who is this *you*? Is the feeling that you make genuine choices real or an illusion?

**Figure 1**
Shall I take a drink?

The mainstream view in science and philosophy is that your feeling of choosing freely is not physically accurate. Just as our intuition of a flat earth has been superseded by a more accurate scientific theory, mainstream science tells us our intuition of having free will has been superseded by a more accurate scientific understanding of neurobiological processes. Like a robot executing a program, our choices are the result of neurobiological processes unfolding according to effectively deterministic physical laws. Pearl and Mackenzie [1] posit that the "illusion of free will" gives us an evolutionary advantage by enabling a compact high-level representation of goals, actions, and priorities. We perform better and learn more efficiently if we use a shorthand representation in which detailed micro-level instructions are encoded in terms of a few high-level options. Indeed, we adopt a similar shorthand when talking about computer programs: rather than describing the details of algorithm execution, we say the loan processing system "decided" to reject an application because of missed car payments, or the robot "chose" to take the longer route to avoid an obstacle.

Are our choices like those of a computer program? Is our experience of making free choices no more than a shorthand encoding of complex but effectively deterministic physical processes in our brains and bodies? This presumption is not as airtight as many assume. Klemm [2] lists twelve major interpretative problems with experiments purporting to support the "zombie argument" that our conscious minds are passive spectators to unconsciously generated actions. Lavazza and De Caro [3] argue that many claims of neural determinism are overstated. On the theoretical side, Stapp [4], [5] argues that genuine, efficacious free will is fully compatible with a realistic interpretation of quantum physics. He argues that there are complementary explanatory gaps in psychology and physics that can be filled by positing an interaction between the brain's quan-

tum state and our mental experience when we make a choice. This interaction provides an opening for efficacious free choice that is fully consistent with the laws of quantum theory [6], [7]. Thus, Stapp argues, genuine free will is fully compatible with our present-day understanding of physics.

The remainder of this article expands on this idea. Section "Free Will Postulates" sets forth properties that formalize a commonsense notion of free will. Section "Free Choice by Physically Embodied Agents" presents a theory of free will that satisfies the postulates given in the Section "Free Will Postulates" and is fully compatible with the laws of physics. Section "Evaluating The Quantum Reducing Agent Hypothesis" describes a path to scientific evaluation of the theory. Finally, implications for science, society, and the future of artificial intelligence are discussed.

## FREE WILL POSTULATES

Arguments about free will often center on whether a notion of free will can be defined that is compatible with effectively deterministic micro-level decision-making processes. Any suggestion that there may be something more to our feeling of free will is dismissed as incompatible with science. The objective of this paper is to formulate a theory of free will that is both compatible with our intuitions and scientifically viable. This requires a clear definition what is meant by the term "free will." The following postulates, taken from [8], are intended to capture and formalize fundamental intuitions of what it means for an agent to have free choice. A similar set of properties is proposed in [9]. An *agent* is defined as a physical system that makes choices satisfying these *free will postulates*:

P1. *Freedom*. There are occasions, called *choice points*, at which multiple alternatives for an agent's future behavior are possible.

P2. *Attribution*. At each choice point, the agent's free choice determines which of the possible alternatives occurs.

P3. *Efficacy*. Choices are efficacious in the sense that the alternative taken at a choice point causes effects in the physical world that are different from what would have occurred had a different alternative been chosen.

P4. *Physicality*. The choices agents make and the effects they have in the world are consistent with the laws of physics.

Many scientists and philosophers take as a given that these properties are mutually inconsistent. P4, the physicality condition, seems to imply determinism, perhaps accompanied by randomness at the quantum level. Determinism violates P1. Randomness violates P2. Compatibilists argue that P1 and P2 should be understood not as actual physical properties, but as our cognitive level experience. That is, our behavior is *actually* mostly deterministic with a bit of quantum randomness, but we *experience* ourselves to be making choices.

In truth, properties P1–P4 are mutually consistent. All four postulates are satisfied by Stapp's [6] theory of free will, which is founded on a realistic interpretation of von Neumann's [10] formulation of quantum theory. The following sections formalize this theory of how mental intentions give rise to bodily actions.

In addition to Properties P1–P4, two additional postulates are needed for a theory of free choice in intelligent, physically embodied agents. Postulates P5–P6 below generalize to arbitrary agents the postulates for human agents presented in [8].

P5. *Representation*. In a manner compatible with their physical architecture, intelligent, physically embodied agents can form representations of the world. They are able to manipulate these representations to predict the effects of the available options and compare the desirability of different options.

P6. *Implementation*. In a manner consistent with their physical architecture, intelligent, physically embodied agents can enact their choices to cause their bodies to behave as intended.

Properties P5 and P6 capture the requirement that intelligent agents are *physical symbol systems* in the sense of Newell and Simon [11]. For biological agents, the physical architecture refers to neurobiological processes in their brains and bodies. For robotic agents, free will would require physical hardware compatible with P5 and P6. What constitutes such a physical architecture is an open scientific question.

## FREE CHOICE BY PHYSICALLY EMBODIED AGENTS

Our experience of having free will is undeniable. We experience our choices as having a causal impact on the world, and *our freely chosen intentions* as the cause of the choices we make. Postulates P1–P4 formalize the intuition behind the notion of free choice. This section shows that these four postulates are consistent with quantum theory and can provide the basis for a theory of agency in nature.

### CAUSAL MARKOV PROCESSES

A scientific theory of agency requires a formal representation of the alternative actions agents can take and how they affect the world. To that end, causal Markov processes provide a formal language for representing the choices of agents and their effects on the environment.

**Definition 1:** *A (time-invariant, first-order, discrete) causal Markov process is a family of stochastic processes specified by the 3-tuple* $(S,A,\pi)$*, where S is a state space, A is an action space, and* $\pi$ *is a transition distribution, such that the following conditions are satisfied:*

1. *For each* $s,s' \in S$ *and* $a \in A$*, the function* $\pi(\cdot|s';a)$ *is a discrete probability measure on S.*
2. *Given an initial state* $s_0$ *and conditional distributions* $\theta(a_k|h_k)$*,* $k = 1,\ldots,n$ *for selecting actions conditional on*

*the past history* $h_k = (a_1,a_2,\ldots,a_{k-1},s_0,s_1,s_2,\ldots,s_{k-1})$ *of actions and states, the joint distribution for the sequence* $(a_1,a_2,\ldots,a_n,s_1,s_2,\ldots,s_n) = (\boldsymbol{a},\boldsymbol{s})$ *of actions and states satisfies*:

$$\mathrm{P}(\boldsymbol{a},\boldsymbol{s} \mid s_0) = \prod_{k=1}^{n} \theta(a_k|h_k)\,\pi(s_k|s_{k-1},a_k). \tag{1}$$

3. *An intervention* $do(a_k = a^*)$ to replace $\theta(a_k|h_k)$ *with the distribution* $1_{[a_k=a^*]}$ *that places probability 1 on* $a^*$*, changes the joint distribution to:*

$$\mathrm{P}(\boldsymbol{a},\boldsymbol{s} \mid s_0) = \pi(s_k|s_{k-1},a^*)1_{[a_k=a^*]}\prod_{j\neq k}^{n}\theta(a_j|h_j,)\pi(s_j|s_{j-1},a_j), \tag{2}$$

Here, the index $k$ represents choice points at which actions may be taken. The actions $a \in A$ represent interventions taken by an agent that affect future evolution of the system. The states $s \in S$ capture all aspects of the agent and the environment relevant to predicting future states and how they are affected by actions. Equation (1), called the causal Markov condition, implies that the most recent past state and the action taken at the next choice point capture all aspects of the world relevant to predicting the next state. Equation (2) formalizes how interventions work. The notation $do(a_k = a^*)$ represents an intervention to set the action at the $k$th choice point to have value $a^*$. The effect of an intervention is to replace the "unperturbed" probability distribution $\theta(a_k|h_k)$ with the distribution $1_{[a_k=a^*]}$ that assigns probability 1 to $a^*$. Interventions satisfy a locality condition: the only effect on the evolution of the system is to set the $k$th action to $a^*$. All other causal mechanisms remain unchanged [12]. The mapping $\theta(a_k|h_k)$ from the history to a probability distribution on the next action $a_k$ is called the agent's policy. The distribution $\pi(s_k|s_{k-1},a_k)$ for the next state conditional on the previous state and the next action is called the transition distribution.

Our theory of free will ascribes the choice of policy $\theta(a_k|h_k)$ to the agent, subject to relevant physical constraints. The transition distribution is ascribed to nature, and in multi-agent problems, the actions of other agents. In other words, Equation (1) specifies how the states of the agent and environment evolve, under the agent's chosen policy, the transition distribution chosen by nature, and the policies chosen by other agents. Equation (2) specifies counterfactual probabilities for the evolution of the agent and environment if the agent were to make different choices.

## QUANTUM THEORY BASICS

Prior to the advent of quantum theory, the evolution of the physical world was thought to be deterministic. Early in the 20th century, this classical picture was definitively overturned by the explicitly probabilistic quantum theory. The formal mathematical foundation for quantum theory was developed by von Neumann [10] in the 1930s. While there is a multitude of ways to interpret the mathematics, von Neumann's formalism remains the standard textbook presentation of quantum theory (e.g., [13], [14]).

The mathematical theory associates a characteristic Hilbert space $\mathcal{H}$ with each quantum system. A Hilbert space is a complex inner product space that is complete with respect to the norm induced by the inner product. The state of a quantum system is represented by a density operator on $\mathcal{H}$, that is, a self-adjoint, positive semidefinite operator with unit trace. A state can be represented as a complex-valued, possibly infinite-dimensional, matrix that is equal to its conjugate transpose, and has real, non-negative diagonal elements that sum to 1. Density operators can represent pure states, statistical ensembles of states, and/or subsystems of a composite quantum system.

A quantum system undergoes two distinct kinds of evolution. The first is continuous, deterministic, mechanical evolution of the quantum state. The second is a stochastic transformation called *reduction*, *measurement*, or more picturesquely, *collapse*.

During mechanical evolution for $d>0$ time units, the initial state $\rho$ transforms to $\mathcal{A}_d\rho$, where $\mathcal{A}_d$ is a completely positive, trace-preserving (CPTP) map that is continuous in $d$ and satisfies $\mathcal{A}_0\rho = \rho$. The CPTP map $\mathcal{A}_d$ depends on the system's environment. For simplicity, the discussion that follows assumes a time-invariant environment, but with appropriate modifications the theory applies to time-varying environments. Reduction corresponds to application of a self-adjoint, bounded operator $R$ to the pre-reduction state $\rho$. The reduction operator can be decomposed as $R = \Sigma_r r P_r$, where the $r$ are real-valued eigenvalues of $R$ and the $P_r$ are mutually orthogonal projection operators summing to the identity. That is, $P_r^2 = P_r$ for each $r$; $P_r P_s = 0$ for $r \neq s$; and $\Sigma_r P_r = I_{\mathcal{H}}$. When a reduction occurs, one of the eigenvalues $r$ is selected with probability $q_r = \mathrm{Tr}(P_r \rho P_r)$, where $\mathrm{Tr}(\cdot)$ denotes the trace. When $r$ is selected, the state instantaneously and discontinuously transforms into the post-reduction state $1/(1/q_r)P_r\rho P_r$.

Quantum theory specifies the rules for evolution between reductions and the probabilities of post-reduction outcomes. However, there is no theory to predict when reductions will occur or which of the allowable reduction operators will be applied. Phenomenologically, reductions have been associated with measurements taken by scientists to observe the system. For this reason, this fundamental gap in quantum theory has been called the "measurement problem".

The mathematics of quantum theory is undisputed, and its probabilistic predictions have been verified to great accuracy. Nevertheless, there has been intense debate over the ontological status of reductions. The many-worlds interpretation asserts that reductions do not actually occur. Instead, each outcome occurs in its own world with its own observers. The question of why we observe only one outcome in our world has not been answered satisfactorily. Realistic interpretations assert that reductions do occur. There have been different proposals to fill the explanatory gap for how and when reductions occur, none of which has gained broad acceptance or achieved empirical confirmation. The Copenhagen interpretation eschews ontological claims, focusing instead on pragmatic rules for predicting the outcomes of experiments.

## QUANTUM THEORY AS A CAUSAL MARKOV PROCESS

The popular conception of quantum theory, with its emphasis on randomness, would seem to provide no room for decision-making. As Searle [15] put it, "It is true that there is an indeterminacy in nature at the quantum level, but that indeterminacy is pure randomness and randomness is not by itself sufficient to give free will."

Yet, randomness is not the whole story. Although it is not widely appreciated, quantum theory can be formulated as an interventionist causal theory [16] with reductions as interventions. As Bohm [17] put it, the quantum state has been called a wave of probability, but it is more accurately described as a "wave from which many related probabilities can be calculated". In other words, quantum theory predicts not a single probability distribution for what will occur, but rather a family of probability distributions, one for each choice of when a reduction occurs and what operator is applied. If the choice of reduction is ascribed to the free will of the agent, this yields a formal theory satisfying properties P1–P4.

Specifically, the *quantum reducing agent hypothesis* (QRAH) postulates that the universe contains systems, called *quantum reducing agents*, that can cause reductions to some parts of their own physical states. According to the QRAH, selection of a policy for initiating reductions is chosen according to the quantum reducing agent's free will. Formally:

**Definition 2:** *A quantum reducing agent (QRA) is a causal Markov process with state space, action space, and transition distribution as given below, where the choice of actions is ascribed to the agent's free will.*

- *State space*: *The states of a QRA are density operators on the Hilbert space $\mathcal{H}$ of the quantum system.*
- *Action space*: *The allowable actions in a QRA are tuples $<d,R>$, where $d$ is a positive real number representing the time until the next reduction and $R$ is a self-adjoint, bounded operator.*
- *Transition distribution*: *Let $\rho$ be the state just after the previous reduction, $d$ the time until the next reduction, $\mathcal{A}_d$ the CPTP map representing mechanical evolution, and $R$ the reduction operator applied after $d$ time units. The initial state $\rho$ evolves mechanically to $\mathcal{A}_d\rho$, at which point the state transitions abruptly to the outcome associated with one of the eigenvalues $r$. The probability of eigenvalue $r$ is given by $q_r = \mathrm{Tr}(P_r\mathcal{A}_d\rho P_r)$. The post-reduction state if $r$ occurs is $\rho_r = (1/q_r)P_r\mathcal{A}_d\rho P_r$. The possible outcomes $\rho_r$ are mutually orthogonal.*

A QRA chooses a policy, or rule for selecting a time at which to initiate the next reduction and an operator to apply. If the system starts at initial state $\rho$ and evolves mechanically for time $d_1+d_2$, the resulting state will be $\mathcal{A}_{d_1+d_2}\rho$. This is the same state that would occur if the no-intervention actions $<d_1, I_{\mathcal{H}}>$ and $<d_2, I_{\mathcal{H}}>$ had been applied to the initial state $\rho$. If the "null" action $<d_1, I_{\mathcal{H}}>$ is replaced by the intervention $do(a_1 = <d_1, R_1>)$, where $R_1$ is a reduction operator, the result is mechanical evolution to $\mathcal{A}_{d_1}$, then a stochastic transition to $(1/q_r)P_r\mathcal{A}_{d_1}\rho P_r$ with probability $q_r = \mathrm{Tr}\left(P_r\mathcal{A}_{d_1}\rho\, P_r\right)$, followed by mechanical evolution to $\mathcal{A}_{d_2}(1/q_r)P_r\mathcal{A}_{d_1}\rho\, P_r$. In general, applying a reduction at any time point causes a stochastic transition at that point, followed by mechanical evolution of the resulting state from that point. This process is illustrated in Figure 2.

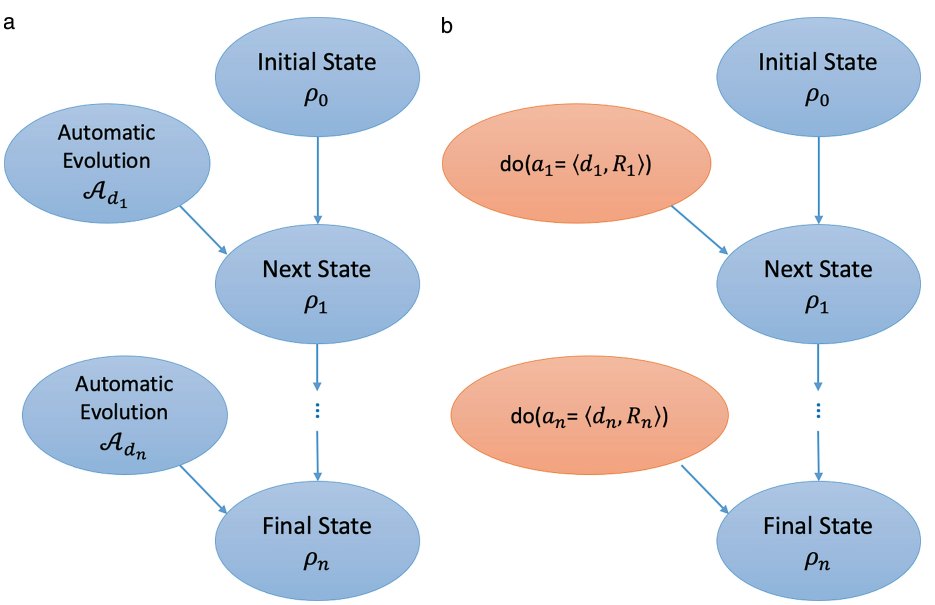


**Figure 2**
Quantum theory as interventionist causal theory: (a) With automatic evolution, $\rho_k$ has value $\mathcal{A}_{d_k}\rho_{k-1}$ with probability 1; (b) On intervention $do(a_k = <d_k, R_k>)$, the state $\rho_{k-1}$ transforms to $\rho_k = (1/q_r)P_r\mathcal{A}_{d_k}\rho_{k-1}P_r$ with probability $q_r = \mathrm{Tr}\left(P_r\mathcal{A}_{d_k}\rho_{k-1}P_r\right)$.

There is a key difference between reductions as typically described in textbooks and the QRAH. Textbooks usually describe measurements a scientist makes on a quantum system undergoing experimental manipulation. That is, the scientist causes a reduction applied to an external system and observes the result, thereby gaining information about the external system. In contrast, the QRAH postulates that a reducing agent causes reductions not directly to an external system, but to *some part of its own physical state*. The two descriptions can be reconciled by recognizing that the scientist's body and the measurement instrument are coupled systems. Thus, if the scientist can effect a reduction to her own physical state that causes her motor cortex to initiate movement of her arm and hand, the hand can then move the control knob on an instrument, which thereby causes a re-

duction of the physical state of the external quantum system. This chain of coupled systems, from the scientist's brain to her body to the external system, is described clearly by von Neumann in his discussion of measurements. The chain ends in subjective perception of the measurement outcome. Subjective perception is, according to von Neumann,

> *...a new entity relative to the physical environment and is not reducible to the latter. Indeed, subjective perception leads us into the intellectual inner life of the individual, which is extra-observational by its very nature. [10], p. 418.*

The quantum reducing agent hypothesis postulates that free will operates via application of state reductions by systems called quantum reducing agents. These agents possess the ability to initiate reductions to some part of their own physical state. They exert free will by choosing which of their available reduction operators to apply at what times. This choice is, to use von Neumann's words, "extra-observational" and is ascribed to the "inner life" of the agent.

The QRAH satisfies postulates P1–P4:

P1. *Freedom*: As currently understood, the laws of physics specify how a quantum system evolves when not subjected to reductions, as well as the probability distribution of outcomes given the reduction operator and time of application. That is, quantum theory specifies the following dynamical laws:

a. $\rho \rightarrow \mathcal{A}_d \rho$ if mechanical evolution occurs for $d$ time units; and

b. $\rho \rightarrow (1/q_r)(P_r \mathcal{A}_d \rho P_r)$ with probability $q_r = \mathrm{Tr}(P_r \mathcal{A}_d \rho P_r)$ if reduction operator $R$ with spectral decomposition $R=\sum_r rP_r$ is applied immediately following mechanical evolution for $d$ time units.

The known laws of physics place no constraints on the choice of time interval $d$ or self-adjoint, bounded operator $R$. Modulo as yet undiscovered limits on $d$ and $R$, there are multiple allowable choices of action $<d,R>$. Therefore, there are multiple possible options at each choice point.

P2. *Attribution*: QRAH attributes the choice of action $<d,R>$ to the reducing agent.

P3. *Efficacy*: The choice of action has empirically distinguishable effects in the physical world, as depicted in Figure 2 and confirmed by extensive experimentation.

P4. *Physicality*: RAH is fully consistent with the known laws of physics as formalized by von Neumann [10] and universally accepted by the scientific community.

By virtue of satisfying P1–P4, QRAH qualifies as a viable candidate theory of efficacious choice by physically embodied agents. We go beyond this basic theory to hypothesize further that QRAs include humans and other life forms, and may also include other kinds of systems in the natural world. In the specific case of human free will, QRAH postulates that human agents make free choices by initiating reductions to some part of their own bodies. Because the cerebral cortex appears to be responsible for cognition and decision-making, it is natural to hypothesize that human QRAs are able to initiate reductions in the cerebral cortex, and specifically in the motor cortex. How, specifically, might such capability be effected in human brains? The next section addresses this question.

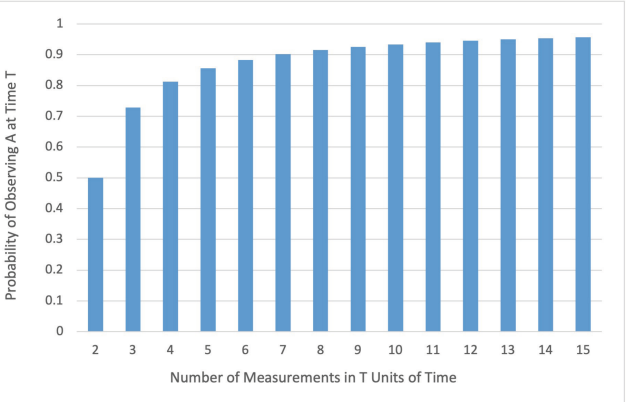


**Figure 3**
Rapid measurement holds quantum system at same state.

## FREE WILL AND THE QUANTUM ZENO EFFECT

Quantum theory textbooks describe reductions as interactions between inanimate microscopic quantum systems and inanimate measuring devices to produce measurements that are observed by scientists. While the founders of quantum theory stressed that the decision of when to initiate a measurement and which measurement to take should be assigned to the free choice of the scientist, they did not consider how such a free choice might be formulated within the scientist's brain and then executed by the scientist's body. In the mathematical formalism of quantum theory, the choice of measurement is free in the specific sense that the theory provides no rules for how the choice is made. Stapp [6] suggests that this gap could be filled by postulating that humans, and possibly other QRAs, have the ability to make free choices by initiating quantum state reductions to some part of their own bodies.

James [18] said, "The essential achievement of the will… is to attend to a difficult object and hold it fast before the mind... Effort of attention is thus the essential phenomenon of will." Stapp suggests that it is this "effort of attention" where quantum theory may play a role. He postulates that Jamesian effort of attention occurs through an essentially quantum phenomenon called the quantum Zeno effect (QZE), whereby rapid repeated reductions applied to a quantum system change its observable behavior [19].

To understand how the quantum Zeno effect works, consider a simple example in which a system is measured and outcome $A$ is observed. The system is then allowed to evolve undisturbed for a period of time, at which point the measurement is repeated. Suppose the system is allowed to evolve for $T$ units of time, and a sequence of measurements, each of which has $A$ as a possible outcome, is taken at equally spaced moments between times 0 and $T$. Figure 3, taken from [7], shows the probability, in a simple example system, of observing $A$ at the last measurement as a function of how many measurements

are taken between 0 and the time of the last measurement. The more measurements are taken, the higher the probability that $A$ will be observed at the last measurement. If enough measurements are taken, the system is effectively frozen in place, and the result $A$ occurs with probability near 1. This effect of freezing a system in place by rapid measurement has been verified in the laboratory [20].

Rapid measurement can also be used to drive a quantum system to a desired state [10]. Because it can be used to accelerate rather than slow down evolution of a quantum system, this phenomenon has been called the inverse quantum Zeno effect [21]. Whereas the quantum Zeno effect involves repeated applications of the *same* reduction operator, the inverse quantum Zeno effect uses a sequence of *different* reduction operators, having a sequence of outcomes along a path from the initial to the desired outcome. If such a sequence of operators is applied in rapid succession, the probability is high that the system will follow this path, resulting in the desired state at the end of the process.

The Section "Quantum Theory as a Causal Markov Process" demonstrated that a QRA as specified in Definition 2 satisfies properties P1–P4. Such a QRA can make choices by initiating reductions to some part of its own body. Stapp further suggests QZE operating in the cerebral cortex as the mechanism by which humans take volitional action. This additional hypothesis requires a model of QZE in the cerebral cortex that is consistent with scientific findings on the neurobiology of volitional action in humans.

A common criticism of quantum theories of mind is that decoherence [22], which occurs when quantum systems interact with their environments, would rapidly destroy quantum effects in the brain. Although environmental decoherence does destroy quantum interference, it is important to note that the quantum Zeno effect survives decoherence [5], and thus could plausibly operate in animal brains. Specifically, a theory of free will in humans must satisfy properties P5 and P6. That is, we require a model of volitional behavior operating through QZE that can be effected in human brains.

## EVALUATING THE QUANTUM REDUCING AGENT HYPOTHESIS

To flesh out this theory, it is necessary to develop a concrete model of how reductions are employed to effect purposive behaviors. Once such a model has been formulated, it must be evaluated empirically. We consider three research thrusts to achieve this objective: simulation, laboratory studies, and hardware implementation. This section is taken, with light edits, from [8].

### SIMULATING A REDUCING AGENT

Stapp proposes, based on research in neuroscience of animal behavior, that the brain learns complex patterns of neurological activity that he calls templates for action. When such an action template is executed, a sequence of nerve signals is sent to the muscles, causing bodily movements. These movements can be adjusted during execution in response to inputs from the senses. The processes of learning and executing these action templates is well described by standard models in neuroscience, e.g., [23]. Where quantum theory plays an essential role, according to Stapp, is to hold an action template in place for longer than it would through purely automatic execution. Thus, the brain automatically retrieves an action template, and QZE is then applied to hold it in place long enough to execute the associated behavior. This model is consistent with the well-established time lag (e.g., Libet [24]) between neural activity associated with a decision and conscious awareness of the decision. An action template is called up prior to conscious awareness of making a decision. The agent then either applies QZE to reinforce the decision or disrupts execution of the action template.

Synchronous oscillations of activity in the brain's neural network appear to play an important role in cognitive processes [25]–[27]. Synchronicity has been hypothesized as a mechanism for how the brain binds component features into representations of composite objects. For example, *in vivo* studies in behaving animals have found that neurons responding to individual features begin firing synchronously when the animal recognizes that the features form a coherent object [28]. Synchronous oscillations also appear to play an important role in motor control [29], preparation for motor activity [30], sensory motor coordination, and focused attention [26]. These findings suggest that templates for action may be characterized by periods of synchronous oscillation in areas of the brain associated with the action to be executed.

Other research suggests a feedback relationship between neural activity and the brain's electrical field [31], [32]. Externally applied electromagnetic (EM) fields have been found in laboratory studies to affect neural activity, and are used in a clinical setting to diagnose and treat a range of neural disorders. Fröhlich and McCormick [31] studied the brain's endogenously generated electric field in a series of *in vivo* experiments and in a computational simulation. Their findings provide evidence of a feedback process in which endogenous electric fields act in a feedback process in which synchronous oscillations increase the strength of the brain's electric field, which in turn reinforces synchronicity of oscillations. Several authors have suggested the brain's electric field as the locus for consciousness (e.g., [4], [33]–[35]). Although the EM field hypothesis is considered speculative, its proponents argue that it explains how information distributed among millions of neurons is unified into coherent percepts. Regardless of the role played by the electric field in consciousness, its role in entraining synchronicity in neural activity appears to be important.

In light of the important role played by oscillations in the brain's electric field, Stapp ([6], Appendix F) developed a simple model of the use of QZE to control the strength of the electric field. His model employed a single frequency quantum oscillator at 20 Hz. The choice of frequency was based on an experimental study that found beta range (15–30 Hz) oscillations in the motor cortex of trained monkeys approximately 100 ms after the monkeys were instructed to move [36]. He also noted that beta oscillations in cortical minicolumns are at the

quantum scale [37], thus suggesting the possible relevance of quantum effects. His single frequency oscillator model can be solved exactly, being a natural extension of the classical simple harmonic oscillator. His analysis demonstrated that the inverse quantum Zeno effect can be applied to increase the amplitude of the quantum oscillator. He calculated the rate of reductions required to have a high probability of increasing the amplitude and concluded that the time scale was reasonable for the neuroscience domain. The amplitude of oscillation corresponds to the strength of the electric field. Thus, Stapp's stylized model demonstrates that the inverse quantum Zeno effect can be applied to increase the strength of the electric field, which in turn would enhance synchronicity in oscillations in the brain's neural network.

Stapp's model considered the oscillating electric field in isolation, without considering how it affects and is affected by synchronicity in neural firing. His model, while suggestive that the QZE could be employed in scenarios consistent with known neuroscience, needs to be extended to a more realistic neurodynamic model.

A potential avenue of research would be to formulate a model that explicitly considers the interaction between the electric field and the spreading activation process in the neural network. The Fröhlich and McCormick model [31] does just this. The model contains some stochastic elements, but is not quantum. Adding quantum effects to a model like this would yield a concrete, biologically plausible model that could be used to investigate the quantum reducing agent hypothesis. Such a model could be used to examine whether rapid reductions can generate macroscopically distinguishable effects on synchronicity of neural activity at biologically realistic parameter settings. The rate of application of state reductions could be included in the model as an adjustable parameter. Reductions could be employed to nudge the brain toward synchronous firing of neurons associated with an action template the organism intends to execute, or to disrupt synchronous firing and thus interrupt an action template the organism intends to discontinue.

Once such a model was developed, it could be implemented on a computer and systematic experimentation could be performed to investigate whether the rate of reduction can be adjusted to entrain or disrupt synchronicity of neural firing. Once neurons are firing synchronously, are there rates of reduction, i.e., "attention density settings", that either reinforce or disrupt synchronous firing? If neurons are not firing synchronously, can "attention density" be employed to generate synchronicity? These and other pertinent questions could be addressed through computational experiments.

It should be noted that the kind of simulation envisioned here should have similar computational complexity to models commonly used in neuroscience. Because environmental decoherence suppresses quantum interference, the quantum neurodynamic model could be approximated as a probability mixture of near-classical possibilities. In other words, extending the approach taken by Stapp in Appendix F of [6], it should be possible to model QZE by modifying a standard stochastic neural network model, thus avoiding the computational difficulties of representing and simulating high-dimensional density operators.

If computational experiments demonstrated that different "attention density settings" produced clearly distinguishable differences in synchronicity using biologically realistic parameter settings, it would lend support to the reducing agent model of efficacious choice.

## LABORATORY STUDIES

Previous sections have articulated a set of hypotheses about how reducing agents may influence the world through the application of QZE. Specifically, the templates for action that guide automatic processing appear to involve waves of synchronous oscillation of relevant parts of the brain's neural network in a feedback relationship with the brain's endogenous electric field. It is hypothesized that QRAs apply QZE to hold desired action templates in place and to apply fine-tuned guidance for their execution. This suggests that rapid reductions would occur in parts of the brain associated with intentional action, and would be employed to increase synchronous firing of neurons associated with action templates the agent intends to implement.

The Section "Causal Markov Processes" proposed developing concrete mathematical models for how QZE influences synchronous firing in neural networks. Such modeling should be informed by laboratory research on the structure and behavior of biological neural networks. Computational experiments with the resulting models could be used to examine the biological plausibility of the hypothesis that efficacious choice operates via the quantum Zeno effect. If successful, these computational experiments should give rise to predictions about the biological mechanisms underlying volition, attention and motor control. These predictions could be tested in laboratory experiments on animals. Results from the laboratory could then be used to refine the computational models and generate additional predictions for further laboratory experiments. The resulting feedback cycle would, if successful, increase our understanding of the neurobiological processes underlying volitional action.

## HARDWARE IMPLEMENTATION

Intelligent agents form representations of the world around them, learn better representations through environmental feedback, manipulate their representations to predict the consequences of different actions, and use these predictions to take intentional action. These representations are formed and manipulated in a physical substrate. Artificial intelligence (AI) has taken the computational metaphor as a given and assumed that the physical substrate of digital computers is sufficient for intelligent behavior. AI has thus pursued the objective of building artificially intelligent agents executing on digital computers.

If the reducing agent hypothesis is correct, then the best that can be hoped for with present-day digital computer systems is a simulation of intelligence. These simulations have performed extremely well on some tasks and less well on others. The reducing agent hypothesis suggests that at least some of the failures may be due to intrinsic limitations of digital computers.

Under the reducing agent hypothesis, achieving true engineered intelligence would require a physical substrate capable of supporting efficacious action through the employment of reductions. That is, an agent's cognitive and motor architecture must be instantiated in a physical structure that can produce macroscopically distinguishable behaviors from different policies for applying reduction operators. The agent must also have a sensory apparatus to convey the real-world results of behavior to a learning system capable of refining the agent's world representation in response to environmental feedback. Research on computational simulations, informed by animal experiments, could inform hypotheses about the kind of physical substrate needed for reducing agents. This research program, if successful, could ultimately lead to engineered intelligence that is more than a simulation.

## DISCUSSION

The stakes in the debate over free will are high. Absent free will, is there any moral basis for expecting socially adaptive behavior or assigning personal responsibility for our actions? Would widespread belief that free will is an illusion lead to nihilism and social dissolution? As Smilansky, quoted in [38], put it, "We cannot afford for people to internalize the truth" about free will. But what if Smilansky's "truth" is not actually true? Is it not critical for science to investigate this question?

It turns out that the four commonsense postulates of freedom, attribution, efficaciousness, and physicality are indeed mutually compatible. All four postulates are satisfied by a realistic interpretation of quantum theory in which physically embodied agents can cause quantum state reductions to some part of their physical states.

Two additional postulates, representation and implementation, must be satisfied by physical symbol systems. Such physical symbol systems might be the "new entity" von Neumann associated with the "intellectual inner life of the individual." That is, causing reductions in the cerebral cortex via the quantum Zeno effect might be the way the "intellectual inner life of the individual" is empowered to make free choices and implement them in the physical world.

The theory presented here is consistent with the known laws of physics, but must be regarded as provisional until it is further fleshed out into a concrete model of behavior in biological systems, and then evaluated empirically. Whatever the ultimate verdict, the profound implications of a physically grounded theory of free will argues for taking the quantum reducing agent hypothesis seriously enough to devise and conduct such tests of its plausibility. Computational experiments could be used to evaluate its consistency with known results in neuroscience. Such experiments could lead to laboratory experiments on animals, and ultimately to a better understanding of decision-making in biological agents.

The implications of the quantum reducing agent hypothesis for the future of artificial intelligence are even more profound. The prevailing view in artificial intelligence is that classical computing theory is an adequate foundation for artificial intelligence. Research in quantum computing focuses on achieving coherent superpositions of many qbits. In contrast, the quantum reducing agent hypothesis suggests that an appropriate physical substrate for engineered intelligence might be an artificial neural network at the edge of the quantum scale that is well-approximated by a classical probability mixture. According to the theory presented here, this is the kind of system that could possess an ability to initiate and control reductions to its own physical state. Computational experiments like those suggested above might give insight on the physical properties required for such a system. This would provide a theoretical basis for intelligent robotic agents with the ability to make genuine free choices.

## REFERENCES


1. Pearl, J., and Mackenzie, D. *The Book of Why: The New Science of Cause and Effect* (1st ed.). New York: Basic Books, 2018.
2. Klemm, W. R. Free will debates: Simple experiments are not so simple. *Advances in Cognitive Psychology*, Vol. 6, (Aug. 2010), 47–65, doi:10.2478/v10053-008-0076-2.
3. Lavazza, A., and De Caro, M. Not so fast. On some bold neuroscientific claims concerning human agency. *Neuroethics*, Vol. 3, 1 (Apr. 2010), 23–41, doi:10.1007/s12152-009-9053-9.
4. Stapp, H. P. *Mind, Matter and Quantum Mechanics* (3rd ed.). Berlin: Springer, 2009.
5. Stapp, H. P. *Mindful Universe: Quantum Mechanics and the Participating Observer* (2nd ed.). Berlin; New York: Springer, 2011.
6. Stapp, H. P. *Quantum Theory and Free Will: How Mental Intentions Translate into Bodily Actions* (1st ed.). New York: Springer, 2017.
7. Laskey, K. B. Acting in the world: A physical model of free choice. *Journal of Cognitive Science*, Vol. 19, 2 (2018), 125–163.
8. Laskey, K. B. A Theory of physically embodied and causally effective agency. *Information*, Vol. 9, 10 (Oct. 2018), 249, doi:10.3390/info9100249.
9. Walter, H. *Neurophilosophy of Free Will: From Libertarian Illusions to a Concept of Natural Autonomy*. Cambridge, MA: MIT Press, 2001.
10. von Neumann, J. *Mathematical Foundations of Quantum Mechanics*. Princeton, NJ: Princeton University Press, 1955.
11. Newell, A., and Simon, H. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, Vol. 19, 3 (1976).
12. Pearl, J. *Causality: models, reasoning, and inference* (2nd ed.). Cambridge, U.K.; New York: Cambridge University Press, 2009.
13. Nielsen, M. A., and Chuang, I. L. *Quantum Computation and Quantum Information*. Cambridge, UK: Cambridge University Press, 2000.
14. Shankar, R. *Principles of Quantum Mechanics*. New York: Plenum, 1994.
15. Searle, J. R. *Freedom and Neurobiology: Reflections on Free Will, Language, and Political Power*. New York: Columbia University Press, 2006.
16. Woodward, J. Causation and Manipulability. In *The Stanford Encyclopedia of Philosophy*, Winter 2016, E. N. Zalta, Ed. Stanford, CA: Metaphysics Research Lab, Stanford University, 2016. https://plato.stanford.edu/archives/win2016/entries/causation-mani/, last access Jul. 06, 2018.
17. Bohm, D. *Quantum Theory*. New York: Prentice-Hall, 1951.
18. James, W. *Psychology: The Briefer Course*. Mineola, N.Y: Dover Publications, 2001.
19. Misra, B., and Sudarshan, E. C. G. The Zeno's paradox in quantum theory. *Journal of Mathematical Physics*, Vol. 18, (Apr. 1977), 756–763, doi:10.1063/1.523304.

20. Patil, Y. S,. Chakram, S., and Vengalattore, M. Measurement-induced localization of an ultracold lattice gas. *Physical Review Letters*, Vol. 115, 14 (Oct. 2015), 140402, doi:10.1103/PhysRevLett.115.140402.
21. Altenmüller, T. P., and Schenzle, A. Dynamics by measurement: Aharonov's inverse quantum Zeno effect. *Physical Review A*, Vol. 48, 1 (Jul. 1993), 70–79.
22. Zurek, W. H. Decoherence and the transition from quantum to classical—Revisited. In *Quantum Decoherence*. Basel, Switzerland: Birkhäuser Basel, 2006, 175–212, doi:10.1007/978-3-7643-7808-0_1.
23. Güçlü, U., and van Gerven, M. A. J. Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in Computational Neuroscience,* Vol. 11, (2017), doi:10.3389/fncom.2017.00007.
24. Libet, B., Gleason, C. A., Wright, E. W., and Pearl, D. K. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain*, Vol. 106, 3 (Sep. 1983), 623–642.
25. Ward, L. M. Synchronous neural oscillations and cognitive processes. *Trends in Cognitive Sciences (Regular Ed.)*, Vol. 7, 12 (Dec. 2003), 553–559.
26. Uhlhaas, P. J. et al. Neural synchrony in cortical networks: history, concept and current status. *Frontiers in Integrative Neuroscience*, Vol. 3, (Jul. 2009), doi:10.3389/neuro.07.017.2009.
27. Wang, X.-J. Neurophysiological and computational principles of cortical rhythms in cognition. *Physiological Reviews*, Vol. 90, 3 (Jul. 2010), 1195–1268, doi:10.1152/physrev.00035.2008.
28. Hirabayashi, T., and Miyashita, Y. Dynamically modulated spike correlation in monkey inferior temporal cortex depending on the feature configuration within a whole object. *Journal of Neuroscience,* Vol. 25, 44 (Nov. 2005), 10299–10307, doi:10.1523/JNEUROSCI.3036-05.2005.
29. van Wijk, B. C. M., Beek, P. J., and Daffertshofer, A. Neural synchrony within the motor system: what have we learned so far? *Frontiers in Human Neuroscience*, Vol. 6, (Sep. 2012), doi:10.3389/fnhum.2012.00252.
30. Tzagarakis, C., West, S., and Pellizzer, G. Brain oscillatory activity during motor preparation: effect of directional uncertainty on beta, but not alpha, frequency band. *Frontiers in Neuroscience*, Vol. 9, (Jul. 2015), doi:10.3389/fnins.2015.00246.
31. Fröhlich, F., and McCormick, D. A. Endogenous electric fields may guide neocortical network activity. *Neuron*, Vol. 67, 1 (Jul. 2010), 129–143, doi:10.1016/j.neuron.2010.06.005.
32. Ye, H., and Steiger, A. Neuron matters: electric activation of neuronal tissue is dependent on the interaction between the neuron and the electric field. *Journal of NeuroEngineering and Rehabilitation*, Vol. 12, (Aug. 2015), doi:10.1186/s12984-015-0061-1.
33. McFadden, J. The CEMI field theory closing the loop. *Journal of Consciousness Studies*, Vol. 20, 1–2 (2013), 1–2.
34. Pockett, S. *The Nature of Consciousness: A Hypothesis*. San Jose, CA: iUniverse, 2000.
35. Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. H. Brain and mind operational architectonics and man-made 'machine' consciousness. *Cognitive Processing*, Vol. 10, 2 (2009), 105–111.
36. Rubino, D., Robbins, K. A., and Hatsopoulos, N. G. Propagating waves mediate information transfer in the motor cortex. *Nature Neuroscience*, Vol. 9, 12 (Dec. 2006), 1549, doi:10.1038/nn1802.
37. Buxhoeveden, D. P., and Casanova, M. F. The minicolumn hypothesis in neuroscience. *Brain*, Vol. 125, 5 (May 2002), 935–951, doi:10.1093/brain/awf110.
38. Cave, S. There's no such thing as free will. *The Atlantic*, (Jun. 2016), https://www.theatlantic.com/magazine/archive/2016/06/theres-no-such-thing-as-free-will/480750/, last access Aug. 22, 2018.



**Kathryn Blackmond Laskey** is Professor of Systems Engineering and Operations Research and Director of the the Center for Resilient and Sustainable Communities (C-RASC) at George Mason University. C-RASC performs transdisciplinary research on sustainable community resilience in the face of disruptions due to technology failures, natural disasters, or malicious human activity. Dr. Laskey's primary areas of expertise are multi-source information fusion, decision support, machine learning, and knowledge representation for reasoning under uncertainty. She has applied her expertise to diverse areas, including crisis response planning, analyzing susceptibility to phishing attacks, detecting insider threats in information systems, predicting innovations in science and technology, protecting soldiers from improvised explosive devices, and understanding airline delays. She teaches courses in systems engineering, Bayesian reasoning, and decision support. She has served on the Board of Directors of the International Society of Information Fusion, was Chair of the Association for Uncertainty in Artificial Intelligence, and currently serves on the Board of Directors of the Washington Metropolitan Area chapter of the International Council on Systems Engineering.