*Interdisciplinary Description of Complex Systems* 13(3), 354-366, 2015

# THE CONCEPT OF FREE WILL AS AN INFINITE METATHEORETIC RECURSION

**Hashim Hanaan[1] and Srikanth Radhakrishna[2, *]**

[1]Sibylgaze, Industry Intelligence GmBH
Bangalore, India

[2]Poornaprajna Institute of Scientific Research
Sadashivnagar, Bangalore, India

DOI: 10.7906/indecs.13.3.2
Regular article

*Received:* 2 April 2015.
*Accepted:* 10 June 2015.

## ABSTRACT

It is argued that the concept of free will, like the concept of truth in formal languages, requires a separation between an object level and a meta-level for being consistently defined. The Jamesian two-stage model, which deconstructs free will into the causally open "free" stage with its closure in the "will" stage, is implicitly a move in this direction. However, to avoid the dilemma of determinism, free will additionally requires an infinite regress of causal meta-stages, making free choice a hypertask. We use this model to define free will of the rationalist-compatibilist type. This is shown to provide a natural three-way distinction between quantum indeterminism, freedom and free will, applicable respectively to artificial intelligence (AI), animal agents and human agents. We propose that the causal hierarchy in our model corresponds to a hierarchy of Turing uncomputability. Possible neurobiological and behavioral tests to demonstrate free will experimentally are suggested. Ramifications of the model for physics, evolutionary biology, neuroscience, neuropathological medicine and moral philosophy are briefly outlined.

## KEY WORDS

free will, uncomputability, infinite recursion, Jamesian two-stage model, quantum indeterminism

## CLASSIFICATION

JEL: I31

PACS: 89.90.+n

*Corresponding author, $\eta$: srik@poornaprajna.org; +91 080 23611836;
Poornaprajna Institute of Scientific Research, 4 Sadashiv nagar,
Bangalore – 560 080, Karnataka, India

## INTRODUCTION

Free will (FW) is a concept in philosophy that refers to the putative capacity of a human agent to control her behavior by choices made by an act of will, on basis of her personal motives, convictions and intentions. The concept rests on the belief that human behavior is not fully determined by external causes. Moreover, her motives, convictions and intentions are themselves not determined by fully external causes, but self-determined. From a common sense perspective, we feel we are free in making decisions. Yet it continues to be debated, even after centuries of argumentation, how to coherently define free will and whether it exists in Nature [1].

FW can be regarded as freedom from some constraint: exactly what that constraint is, has remained moot. There are two broad philosophical positions on FW: *Incompatibilism*, which holds that the relevant constraint is determinism, and *Compatibilism*, which holds that determinism is irrelevant to the definition of free will, and that determinism and FW are compatible.

Two divergent Incompatibilist positions are *Hard Determinism*, which regards FW as false and determinism as true, and (metaphysical) *Libertarianism*, which regards determinism as false and FW as true. A FW *Skeptic* is an Incompatibilist who goes farther than a Hard Determinist and denies that FW is even a coherent concept. From the Skeptic perspective, indeterminism no more allows an agent control and self-determination over actions than does determinism. Since determinism and indeterminism are the only two logically possible causal primitives, this view holds that very concept of FW is meaningless. This Skeptic stand is sometimes called the dilemma of determinism or the standard modern argument against FW.

The libertarian may imagine that housed in her brain is an immaterial agency (such as a soul or homunculus) that somehow transcends the cause and effect law that holds elsewhere in Nature. This is a logically tenable defense of Libertarianism against Hard Determinism, even if an adherent of the latter would deny the existence of such immaterial agencies. But even the soul offers Libertarianism no protection against the Skeptic, because if the soul's choices transcend causality, they must be random, or results of random properties, again making the case for FW bleak!

The present work hopes to convince the FW Skeptic that a coherent metaphysical account of FW is possible. Of importance to our work is using an idea similar to that used by Tarski to define the concept of truth, except that here it is applied to causality. We think that FW is a causal primitive different from both determinism and indeterminism in being *metatheoretic*. Our attempt to define it shows it to be a form of causation that straddles endless levels of causality, *provided* we wish to form a coherent scientific narrative that fits in with out intuitive sense of self, freedom and responsibility.

This article is structured as follows. A version of the Jamesian two-stage model is proposed in Section 2, which would be useful for our further discussion. Although such a model has been considered as the defense of FW against the dilemma of determinism, we point out in Section 3 that the model fails, because the dilemma can be recursively resurrected. In Section 4, we show that an infinite recursion of Jamesian two-stage models restores a measure of protection for libertarianism against the dilemma. This model is adapted to define a Rationalist-Compatibilist FW. The connection of FW to uncomputability [2] and Tarskian undefinability for formal truth [3] are outlined in Section 5. We then discuss the consequences of the model for neurobiology in Section 6, before concluding in Section 7.

## THE TWO-STAGE MODEL, 2*S*: A NEW HOPE

The two-stage model [4], introduced in its original form by W. James [5], and proposed in various forms by a number of researchers, is intended to defeat the dilemma of determinism.

The model posits that FW is a two-stage process: first there is freedom at the lower stage, and then there is will at the higher stage, which makes a choice. We present further in the text a version of it, which we call '2*S*', with several changes in the details, as described in the following two subsections.

## PHYSICAL STAGE *L*0 (FREEDOM)

At the moment an agent's attention is drawn to a conflict situation, during a short time span, called the *selection window*, in a localized region of brain, which we refer to as the *free-will oracle*, probably in the pre-frontal cortex, the physical laws $W_0$ are put on a "causally open" mode. In preparation of entry into this mode, alternative options $x_0$, described by probability distribution $P_0$, are generated in the agent's brain. The physical laws $W_0$ only determine $P_0$, but do not entirely fix the eventual choice $x_0$. And the selection is not completely determined by the past history of the *physical* universe. At the close of the selection window, $W_0$ in FW oracle is re-set to the "causally closed" mode, and the choice $x_0$ that is available on the FW oracle's "register" at that moment is expressed as the agent's action.

If the agent makes the choice mechanically, without a focused exertion of her will, then there is no mental causation influencing the choice (as described below), and at the close of the selection window, a random $x_0$ is selected according to probability distribution $P_0$.

## METAPHYSICAL STAGE *L*1 (WILL)

If the agent decides to exert her will, then her choice must not be random but instead reflect the desires, intentions and beliefs characteristic of the agent. These properties constitute her *cognitive private space*, and may not have a full representation in the physical level $L_0$. Thus the will produces generally a *deviation* from $P_0$.

A simple way to describe this situation mathematically is by:

$$x_0 = \lambda_1(P_0), \tag{1}$$

where $\lambda_1$ is the "will function" that encodes the properties of her cognitive private space. Function $\lambda_1$ takes note of $P_0$, but is *not* part of $W_0$, because if it were, then there would be causal closure in the physical, and hence no freedom at $L_0$. Instead, $\lambda_1$ is determined by the laws $W_1$, which extend $W_0$ to $L_1$.

The fact that the causal openness at $W_0$ is replaced by causal closure of $W_1$ in the consolidated system in $L_0 + L_1$ suggests that these two levels can be ordered in a *causal hierarchy*, which we represent by the expression: $L_1 < L_0$. The ordering expresses that $L_1$ causally precedes $L_0$.

## 2*S* AND THE DILEMMA OF DETERMINISM

In the absence of action by the will, the FW oracle can be considered as a *probabilistic* input/output machine. To underscore this, sometimes we will refer to such indeterministic behavior as *freedom-without-will*. We introduce a level $L_{0.5}$ as essentially $L_0$ equipped with a source of pure randomness. Instances of freedom-without-will have their causal closure in $L_0 + L_{0,5}$, but in this case there is no deviation from $P_0$. When the will is exerted, the random variable $X_0$ representing free choice deviates from $P_0$, and free will transcends physical causality.

The attempted defense of FW against the dilemma of determinism using model 2*S* would be: FW is not deterministic because of freedom at the physical stage. Nor is it random because the eventual choice is self-determined, being fixed by personal preferences via $\lambda_1$.

Though not the definitive word on FW, the model 2*S* brings the new insight that FW is a new *causal primitive*, different from determinism and indeterminism. While the latter two can be associated with a causally closed lawfulness, and thus defined on a single causal level (say $L_0$), FW cannot. More of this further in the text.

### FREE WILL IN 2*S* AND FORMAL TRUTH

As we indicated in Ref. [6], the 2*S* feature of introducing the metaphysical level to define FW on the physical level parallels Tarski's use [3] of a metalanguage in order to consistently define the concept of arithmetic truth in an object language. Tarski showed that without careful separation of the two levels, one would end up with logical antinomies like the liar's paradox "This statement is false" (which is true iff it is false).

Similarly, if we fail to separate the "free" and "will" stages in free choice, we will err in reducing the choice to determinism or randomness. We believe that part of the difficulty in understanding the nature of FW is due to a lack of appreciation for this level separation.

With this in mind, we may consider $L_0$ as the object stage, and $L_1$ as the metastage. The free-willed agent with 2*S* structure will be designated $\boldsymbol{F}_1$. A deterministic physical system, which is causally closed at level $L_0$ is designated $\boldsymbol{F}_0$, while an indeterministic quantum physical system, which has $L_0$ freedom-without-will will be denoted $\boldsymbol{F}_{0,5}$.

## DILEMMA OF DETERMINISM REVISITED

Although the model 2*S* resolves the dilemma of determinism after a fashion, still the dilemma can be resurrected at level $L_1$. To see this, note that the two-stage agent $\boldsymbol{F}_1$, taken as a whole, is deterministic, in view of Eq. (1), and must thus lack FW. This is just Schopenhauer's argument, who picturesquely said, "Man can do what he wants but he cannot will what he wants" [7], in his prize-winning essay in response to the challenge posed by the Royal Norwegian Academy of Sciences in 1839[1].

To reinstate free will, we allow the will function $\lambda_1$ to be freely chosen, i.e., we apply 2*S* to the selection of $\lambda_1$. We extend the FW oracle from the physical level $L_0$ to the metaphysical $L_1$. Within the selection window, the laws $W_1$ governing level $L_0 + L_1$ become momentarily causally open in the region of the FW oracle. We thus have *freedom of the will* at $L_1$ whereby the $L_0 + L_1$ (i.e., physical-metaphysical) history of the universe does not fix $\lambda_1$.

We introduce a metalevel $L_2$, where a higher-order will function $\lambda_2$ ("will free will"), provides causal closure, by deterministically selecting a particular $\lambda_1$. We can think of $\lambda_2$ as representing a deeper aspect of the agent's character or disposition and as determining the type of the will function $\lambda_1$ that she will select. Function $\lambda_2$ comes from a further interior layer of the cognitive private space of the agent. By adding an element of spontaneity and self-determination, it thwarts the selection of $\lambda_1$ from being modelled as an input/output process at level $L_1$. In the metatheoretic representation, this is tantamount to treating $\boldsymbol{F}_1$ as the object system, and an $L_2$-aspect as the metasystem. A simple mathematical way to represent this is by:

$$\lambda_1 = \lambda_2(P_1), \tag{2}$$

where $P_1$ is a *probability function* encoding $L_1$-preferences according to laws $W_1$. If the higher-order will $\lambda_2$ is not exerted, then $P_1$ would describe the indeterministic selection of will $\lambda_1$. But if this "will free will" is exerted, then there will be deviations from $P_1$ in the selection of $\lambda_1$. Substituting Eq. (2) into Eq. (1), we obtain:

[redacted] (3)

The second-order will $\lambda_2$ selects a first-order will $\lambda_1$ depending on $P_1$, and then $\lambda_1$ selects $x_0$. The extended causal hierarchy is $L_2 < L_1 < L_0$, whereby $L_2$ causally precedes $L_1$, which in turn precedes $L_0$.

Here precedence refers not just to chronology but to recursion depth in the agent's cognitive private space. It does not refer to an *earlier* event on the same causal level, but to *deeper* or higher-order cause. For example, if $\lambda_1$ inclines Alice to give Bob a gift, then $\lambda_2$ could be her character trait of wanting to help him. A yet deeper cause ($\lambda_3$) would be wanting for good to him or people she likes. And so on.

We denote by $\boldsymbol{F}_2$ the type of agent for who there is causal closure at $W_2$, and by $\boldsymbol{F}_{1,5}$ one for who the will function $\lambda_1$ is selected indeterministically. Although the existence of $\boldsymbol{F}_2$ resolves the dilemma of determinism on $L_1$, still the dilemma can be resurrected at level $L_2$. This is evident seeing that the action of $\lambda_2^*$ in Eq. (3) is deterministic.

To reinstate free will at $L_2$, we recursively apply 2$S$. We make $L_2$-law $W_2$ at the FW oracle also causally open within the selection window. Let the freedom of $\lambda_2$ be described by probability function $P_2$, determined by $W_2$. Causal closure occurs at $W_3$ via higher-order FW $\lambda_3$ from an even deeper aspect ($L_3$-aspect) of the agent's cognitive private space. We have

$$\lambda_2 = \lambda_3(P_2), \tag{4}$$

Substituting Eq. (4) into Eq. (2), we obtain:

$$x_0 = [\lambda_2(P_1)](P_0) = [[\lambda_3(P_2)](P_1)](P_0) \equiv \lambda_3^*(P_2, P_1, P_0). \tag{5}$$

We extend the causal hierarchy as $L_3 < L_2 < L_1 < L_0$. An agent with causal closure of her choosing process in $W_3$ is denoted $\boldsymbol{F}_3$ and one with freedom-without-will on $L_2$ (i.e., $L_2$-indeterministic) is denoted $\boldsymbol{F}_{2,5}$.

Although the existence of $\boldsymbol{F}_3$ appears to resolve the dilemma of determinism on stage $L_2$, still the dilemma can be resurrected at stage $L_3$ since $\boldsymbol{F}_3$, in view of Eq. (5), can be considered as a deterministic system. This may be seen as a further higher-order extension of Schopenhauer's argument.

To prevent the dilemma of determinism at stage $L_3$, we introduce a 2$S$ model on top of this stage, with $\boldsymbol{F}_3$ as the object system, and $L_4$ as the metastage. But then the resultant $\boldsymbol{F}_4$ will be deterministic. We require $L_5$ to obtain a 2$S$ model for $\boldsymbol{F}_4$ to avoid the dilemma of determinism at level $L_4$. Continuing this trend indefinitely, at level $L_n$, where $n$ is any positive integer, we have:

$$\lambda_{n-1} = \lambda_n(P_{n-1}). \tag{6}$$

Substituting this recursively into lower levels into Eq. (3), we obtain:

$$x_0 = [[\cdots[[\lambda_n(P_{n-1})](P_{n-2})]\cdots](P_1)](P_0) \equiv \lambda_n^*(P_{n-1}, \cdots, P_0). \tag{7}$$

Evidently, we can still resurrect the dilemma of determinism at level $L_{n+1}$ no matter how large $n$ is, since the fact of $\lambda_n^*$ in Eq. (7) is a deterministic function. Thus, the problem posed by this dilemma does not disappear for agent $\boldsymbol{F}_n$ but is merely postponed.

In response to this seemingly insurmountable difficulty posed by the recursive version of the dilemma of determinism, the only option to save libertarianism seems to be to let a free-willed agent be an infinite-stage entity, $\boldsymbol{F}_\infty$. What this means is that, for any *finite* integer $n$, the laws ($W_n$) of cause and effect at the stage $L_n$ will lack causal closure in the region of the extended FW oracle during the selection window, and the selection of will $\lambda_n$ cannot be modelled as a probabilistic or deterministic input/output system at stage. The causal closure for $W_n$ will come through a higher-order cause $\lambda_{n+1}$, which is determined by the $L_{n+1}$-aspect of the cognitive private space. Therefore, the agent's choice of $\lambda_n$ will transcend $n^{\text{th}}$-order causality $W_n$, so that the choice of $\lambda_n$ should be regarded as spontaneous and self-determined at that stage.

In FW so understood, there is libertarian freedom in the sense that the agent's free choice has an inexhaustible causal depth in her cognitive private space. Perhaps, when we scan the inner space of our consciousness, and feel that our choices are free, it is this infinitude that we grasp intuitively, and feel inclined to report as genuine personal freedom.

Identifying human agents with $\boldsymbol{F}_\infty$ also means that FW is at least a *supertask* [8], a process that involves a sequence of countably infinite number of steps executed in finite time. In Section 5, we will present an argument suggesting that free choice is probably even a *hypertask*, which involves uncountably many steps executed in finite time.

# FREE WILL AS AN INFINITE METATHEORETIC RECURSION

The free agent $\boldsymbol{F}_\infty$ in our model is an infinite-stage entity straddling the physical $L_0$ and the "final" or "infinite-th" stage, denoted $L_\infty$. For simplicity, we will refer to the agents $L_0$-aspect as the "physical aspect" and the $L_\infty$-aspect as "transfinite aspect". In a conflict situation, a response is initiated at the transfinite aspect and transmitted to the physical aspect, where it manifests as the choice $x_0$. It stands to reason that the final desination to which information about the sensory input is taken, before the agent's response is initiated, must also be the transfinite aspect.

## AN INFINITE, STAGED CAUSATION

Extending (7), we can represent the choice $x_0$ through a sequence of downward causations starting from the "transfinite preference" $P_\infty$:

$$x_0 = [[[\cdots[[[\lambda_\infty(P_\infty)]]\cdots]](P_2)](P_1)](P_0) \equiv \lambda^{\#}_n\,[P_n, P_{n-1},\cdots, P_\infty](P_{n-1},\cdots, P_0). \tag{8}$$

The interpretation is that $\lambda^{\#}_n$, the will at stage $L_n$, is fixed by higher-order preferences, and then selects an outcome $x_0$ depending on lower-order preferences.

The form of Eq. (8) suggests that the larger the recursion depth $n$, the fewer the higher-order preferences $P_{n+1}$, $P_{n+2},\cdots$ that could sway $\lambda^{\#}_n$ from the motivation encoded by $P_\infty$. In Eq. (8), suppose that $0_{n-1},\cdots,0_0$ represent the probability functions $P_{n-1}$, $\cdots$, $P_0$ that are *unbiased* in the sense of being consistent with $P_\infty$.

Replacing the lower-order preferences by their unbiased values in Eq. (8), we now define the $n^{\text{th}}$-order *intent*

$$x_n = \lambda^{\#}_n\,[P_n, P_{n+1},\cdots, P_\infty](0_{n-1},\cdots, 0_0), \tag{9}$$

meaning that $x_n$ is the choice $x_0$ that *would* be made if there are no distortions downwards from level $L_{n-1}$. Thus we may call $x_\infty$ as the "prime intent" or "transfinite intent", the option that would be selected if the will at infinity, $\lambda_\infty$, were to act unthwarted on the physical.

During the act of free choice, the prime intent is replaced stage-wise by lowerorder intents, until the final choice is reached. We may refer to this infinite train

$$\boldsymbol{X} \equiv x_\infty \to \cdots x_n \to x_{n-1} \to \cdots \to x_1 \to x_0, \tag{10}$$

as the "descent of the will". This immediately evokes the notion of FW as the effectiveness of communication of $x_\infty$ from the transfinite aspect to the physical aspect, undistorted by lower-order preferences. The will is stronger, if this channel of communication ("volition channel") is clearer, uncluttered by lower-order motivations, beliefs and desires inconsistent with their transfinite counterparts. This line of thought forms the basis of the compatibilist FW introduced further in the text.

## RATIONALIST-COMPATIBILIST FREE WILL

The model above can be extended to protect FW from what may be called the 'rationalist/robot paradox'. By definition, a free, rational agent will, when faced with a choice, select the optimal option. His behavior is completely predictable, assuming that there is a single rational option. For a libertarian, rationality appears to undermine freedom [9]. Now this is not the case, as viewed by a compatibilist. But the rationalist/robot paradox asks how the compatibilist would differentiate a rational agent from an optimal robot programmed to choose rationally.

Following the line of thought indicated previously, we would like to think of the correlation between the prime intent and final choice as a measure of FW, since it expresses how well the agent is able to hold on to her prime intent by overcoming deviating influences. However,

this correlation would stay as merely incidental, unless the physical aspect holds $x_\infty$ as her purpose or motive. For this, she must have cognizance of $x_\infty$. Precisely this defines the role played by the agent's rational faculty or reasoning in FW. We will refer to an $\boldsymbol{F}_\infty$ agent equipped with the rational faculty by $\#\mathbf{F}_\infty$.

The role played by reason is crucial. Without it, the physical aspect has no motive to deviate behavior from that determined by physical causality $W_0$. Now there may be random deviations from $P_0$ (applicable to animal agents), but they would be devoid of any systematic or deliberate attempt to transcend $W_0$. By contrast, a human agent, on the recommendation of reason, tries to overcome the imposition of $W_0$ by trying to deviate $X_0$ towards $X_\infty$. Here the quantity $X_j$ represents the random variable corresponding to $x_j$, i.e., values of variables associated with a probability distribution. Thus the reasoning faculty serves as the basis through which the opportunity provided by causal openness is exploited.

An agent is free to the extent that she is able to enforce her transfinite will on her physical choice. (Complications arising from the corruption of the rational faculty will be ignored here.) This gives us a quantification of Rationalist-compatibilist FW:

$$\boldsymbol{G} = \mathrm{Corr}(X_\infty : X_0), \tag{11}$$

where Corr is any measure normalized so that $-1 \leq \boldsymbol{G} \leq +1$. The rational free-willed agent is characterized by $\boldsymbol{G} = 1$, while a person completely under the sway of material nature, by $\boldsymbol{G} = -1$.

Lacking (substantial) reasoning, an animal may be represented simply by $\mathbf{F}_\infty$. The animal is free, but not free-willed. We express this insight with the expression:

$$\text{Freedom + reason = free will} \tag{12}$$

Quantum matter, or in particular quantum AI, which remains under the scope of physical causality, is a $\boldsymbol{F}_{0,5}$ agent, while a classical robot is a $\boldsymbol{F}_0$ agent. It is clear how this Rationalist-compatibilist account protects FW from rationalist/robot paradox: a deterministic robot is a $\boldsymbol{F}_0$ agent, while a rational free-willed agent is a $\#\boldsymbol{F}_\infty$ agent with $\mathbf{G} = 1$.

The brain is arguably a special organ, whose physical structure has somehow been evolved equipping it with a FW oracle, providing a gateway to the transfinite aspect. AI lacks this and the physical laws governing its dynamics are causally closed.

It seems to be an interesting proposition that plants and "lower animals" (like microbes), which lack a central nervous system (CNS), could be considered as intermediary agents between quantum matter and higher animals (like mammals, reptiles and birds, which have a CNS), and thus represented by $\boldsymbol{F}_K$, where $0{,}5 < K < \infty$. Some of these ideas are summarized in Table 1.

**Table 1.** Freedom gives spontaneity, reason gives self-determination, and freedom with reason is free will. AI, lacking a FW oracle and being thus just a special configuration of quantum matter, is described by first-order indeterminism. By contrast, the higher animal or human brain, being equipped with a FW oracle, has freedom at all orders. Perhaps the $\boldsymbol{F}_\infty$ structure, common to humans and animals, is necessary for emotional behavior.

| Entity | Agent type | Resource |
|---|---|---|
| Human | $\#\boldsymbol{F}_\infty$ | Free will (Freedom at all orders, plus reason) |
| Higher Animals (having a CNS) | $\boldsymbol{F}_\infty$ | Freedom at all orders |
| Lower Animals and plants (Lacking a CNS) | $\boldsymbol{F}_K$ $(0{,}5 < K < \infty)$ | Freedom up to a finite order |
| Quantum AI | $\boldsymbol{F}_{0,5}$ | First-order freedom-without-will |
| Classical AI | $\boldsymbol{F}_0$ | Determinism |

### CAUSAL VS. LOGICAL DETERMINISM

The resolution of the rationalist/robot paradox shows that the predictability of behavior does not imply that the behavior was causally determined (as in the robot's case). There is logical determinateness about the rational agent's behavior even though he transcends physical causality. We express this idea by:

$$\text{Logical determinism} \Rightarrow \text{Causal determinism.} \tag{13}$$

Now this result would be undermined if all humans were perfect ($\boldsymbol{G} = 1$) then, even if $P_\infty \neq P_0$, we would be led to suspect that there is a "law of goodness", characterized by $P_\infty$, that controls human behavior. However, some people are imperfect (having $\boldsymbol{G} < 1$), suggesting that human behavior in general transcends causal determinism, and weakening the need to undermine the above conclusion.

### EXPERIMENTAL TEST

It is an interesting and old[1] question: how to experimentally demonstrate the existence of FW? Our model suggests that unfocussed or casual acts of choice would be governed by a probability distribution $P_0$, while a free-willed action with deliberate intent will in general produce a deviation from $P_0$ towards $P_\infty$.

The observation of discrepancy between the statistics of focused and unfocused choice, could be one way to demonstrate the existence of FW. Designing such an experiment may not be easy, since the very act of focusing may psychologically alter $P_0$, so that an observed deviation may either be due to will-induced deviation or due to an alteration of $P_0$ or due to both.

### WHY WE ARE NATURAL LIBERTARIANS

The model also helps make sense of people's instinctive inclination to Libertarianism. It is a reasonable assumption that as an agent introspectively scans the inner space of her consciousness, depending on how subtle her awareness is, she can at best objectively perceive only so deep as there is freedom.

The unfreedom and higher-order causal influences lying beyond that point become part of her subjective consciousness, and she is unable to consciously experience them, though she may deduce them by observing her conscious choices and preferences.

## FREE CHOICE AND UNCOMPUTABILITY

We now explain a line of thought indicated in Ref. [6], on the correspondence between the causal hierarchy and the hierarchy of Turing uncomputability, in reversed ordering. That is, given a $\#\boldsymbol{F}_\infty$ agent, $x_\infty$ is computable, while $x_m$ is harder to compute than $x_n$ if $m < n$.

The basic idea behind this claim is the following. Suppose one has a computer program (or Turing machine) so powerful that it can compute the free choice of a free-willed agent, using the current most detailed description of her brain state. Now if its prediction were shown to the agent, being free-willed, she may contradict the prediction. The conclusion is that such a powerful computer program does not exist. We now consider a somewhat more detailed argument.

Given a free-willed $\#\boldsymbol{F}_\infty$ agent **A**, suppose there is a computer $\mathbf{T}_C$, programmable in some computer language $\mathbf{T}_L$ and suitable for the task of computing **A**'s free choice $x_0$. Let $\hat{\mathrm{A}}$ denote the description of **A** as a computer program in $\mathbf{T}_L$. We assume that all computer programs that encode in $\mathbf{T}_L$ the description of free-willed agents are denumerable and that $\hat{\mathrm{A}}$ is the $a^{\text{th}}$ program. Similarly, one is assumed to be able to encode situation any conflict situation **J** in the medium of $\mathbf{T}_L$ by a description $\hat{\mathrm{J}}$, and enumerate them alphabetically as some number $j$.

If free choice is computable, then the computer $\mathbf{T}_C$ can, given the enumerations for the $\mathbf{T}_L$-description of the agent and the conflict situation, compute the agent's free choice in finite time, or:

$$\tau_C(a;j) = \begin{cases} 0 \Leftrightarrow \mathbf{A}(\mathbf{J}) = 0, \\ 1 \Leftrightarrow \mathbf{A}(\mathbf{J}) = 1. \end{cases}, \tag{14}$$

where for simplicity we have assumed the outcome to be two-valued. (There is no loss of generality, since any computable output can be made binary, for example by assigning "0" if the outcome is non-numerical or numerical and less than 0, and "1" otherwise). In words, the computer produces output 1 (resp., 0) if **A** freely chooses 1 (resp., 0) when faced with conflict situation **J**.

For any positive integer $j$, using this as a subroutine, we can build another program:

$$\tau_R(j) = \begin{cases} 1 \Leftrightarrow \tau_C(j;j) = 0, \\ 0 \Leftrightarrow \tau_C(j;j) = 1. \end{cases} \tag{15}$$

which is a representation of the above notion of the uncooperative free-willed agent. Thus $\mathbf{T}_R$ outputs "0" on input $j$ iff the $j^{\text{th}}$ free-willed agent outputs "1" on the $j^{\text{th}}$ conflict situation.

We can now apply the computer to $\mathbf{T}_R$, so that from Eqs. (14) and (15):

$$\tau_C(t_R;j) = \begin{cases} 1 \Leftrightarrow \tau_C(j;j) = 0, \\ 0 \Leftrightarrow \tau_C(j;j) = 1. \end{cases}, \tag{16}$$

where $t_R$ is the enumeration of $\mathbf{T}_R$. If we set $j := t_R$, then we are led to a contradiction as it would entail that $\mathbf{T}_C(t_R, t_R) = 0$ if and only if $\mathbf{T}_C(t_R, t_R) = 1$.

To restore consistency, we infer that $\mathbf{T}_C$ will never halt on inputs $(t_R, t_R)$, which thereby constitutes an undecidable Gödel sentence under the above encoding. We conclude that in general $x_0$ will be uncomputable for the family of computer programs considered.

One can conceive a higher-level "meta-computer" program $\mathbf{T}_C[(1)]$ that is able to decide whether $\mathbf{T}_R(t_R)$ equals 0 or 1, but that is not contained in this family. If the cardinality of such meta-computers is greater than $\aleph_0$ (countable infinity), then the above diagonal argument based paradox can be averted, because the meta-computers will not be denumerable. This situation is similar to that pertaining to the concept truth, of requiring a metalanguage in order to define truth in the object language [3].

The proof given above for the uncomputability of $x_0$ is similar to the that of the uncomputability of the halting problem for Turing machines [2]. The concept of a meta-computer alluded to above indicates that the proof of uncomputability "relativizes", meaning that one can construct harder problems, by allowing a computer program to call as subroutine an "oracle" that solves the above free choice problem in bounded time. One can then construct a Gödel sentence for this oracle-enhanced program, which then yields a problem with its hardness shifted one level higher than the free choice problem above. Problems on the same level of uncomputability, i.e., uncomputable problems which are Turing-equivalent, form a Turing degree. The process can be repeated to construct higher Turing degrees, i.e., the next higher levels of more uncomputable problem [10]. It is known that there are $2^{\aleph_0}$ (uncountably many) Turing degrees.

We suggest that the the causal hierarchy corresponds to Turing degrees, but in inverse ordering, whereby the prime intent, which arises beyond an infinite number of causal stages, is itself Turing-computable. But the descent of the will would correspond to transition to higher levels of Turing uncomputability, making the free choice of agents of sufficiently low FW highly uncomputable. The act of FW in general must be a monstrous hypertask, since the lower causal stages correspond to ever higher orders of uncomputability.

Why should the causal and computational hierarchy correspond with each other? Here we will appeal to a teleological argument: that if the consequences of the causal hierarchy were computable, then there would have been no need for the "brute force" computation provided by the physical manifestation of the universe and human agents!

Considerable research has been devoted in computability theory and mathematical logic to the study of the complicated structure of Turing degrees. Perhaps all of that may have a bearing on the cognitive structure of free-willed agents.

## NEUROLOGICAL BASIS FOR FW

The presence of the FW oracle in the human brain marks the basic difference between a human agent and a robotic simulation. The question of how the FW oracle is embedded in the brain and called forth, is briefly considered here.

A $\#\boldsymbol{F}_\infty$ agent is an infinite entity, whereas the physical brain of a human being is finite in terms of its information storage and computation capacity. Therefore, if humans are $\#\boldsymbol{F}_\infty$ agents, sufficiently high levels $L_j$ cannot have a physical representation, i.e., a neural correlate.

We propose the following physical realization of the model. At the instance a human becomes aware of a choice, she is instinctively driven to enact her "nature", encoded by $P_0$. At the selection window, her "reason", which carries a representation of $P_\infty$, advices her to deviate $X_0$ towards $X_\infty$. This creates a potential tension, which may result in a fleeting quantum superposition, and may correlate with the agent subjectively experiencing an internal conflict.

Since $P_0$ is determined by $W_0$, its neural correlate is expected to be well defined, and associated with the motor cortex. Since $P_\infty$ is largely computable, its neural correlate is also expected to be well defined, and associated with the reasoning circuits in the pre-frontal cortex. In Figure 1, these two correlates are represented as the slow and fast neural pathways, at whose confluence the FW oracle lies.

From physically observable data, one may be able to predict a pattern of behavior. However, since the higher levels in the causal hierarchy are not observable, therefore in any given instance of an agent's free choice, even the most detailed neural imaging (say via fMRI) will be unable in principle to predict with full certainty what the agent will select.

The FW mechanism in the brain of a $\#\boldsymbol{F}_\infty$ agent can never be modelled as a finite input-output device. An important physical consequence is that, the mental causation that produces a deviation from $W_0$ may correspond to deviations from physical laws associated with $W_0$, like energy conservation or the Second Law of thermodynamics. For example, the initial voltage fluctuation in a motor neuron that initiates the spontaneous movement of a mouse's whisker, may be energy-wise unaccounted physically, even though the subsequent nonlinear amplification of that fluctuation to a physical action will certainly be governed by (classical) physics.

A possible experimental test of the model could aim to distinguish willful from casual choice. In the latter case, the probability of choice will reflect $P_0$, which can be estimated from the relative strengths of signals, as picked up by fMRI scans. Under willful choice, there will be a deviation in the probability of choice away from $P_0$ towards $P_\infty$.

## DISCUSSIONS AND CONCLUSIONS

We showed that, although metaphysical libertarianism is provisionally protected against the dilemma of determinism by the two-stage model, it is vulnerable to the recursive version of the dilemma. As a defense for libertarian freedom, we proposed the infinitely recursive two-stage scheme. A free agent is described here by $\boldsymbol{F}_\infty$, and free-willed agents by $\#\boldsymbol{F}_\infty$. The concept of FW here is of the Rationalist-compatibilist kind.

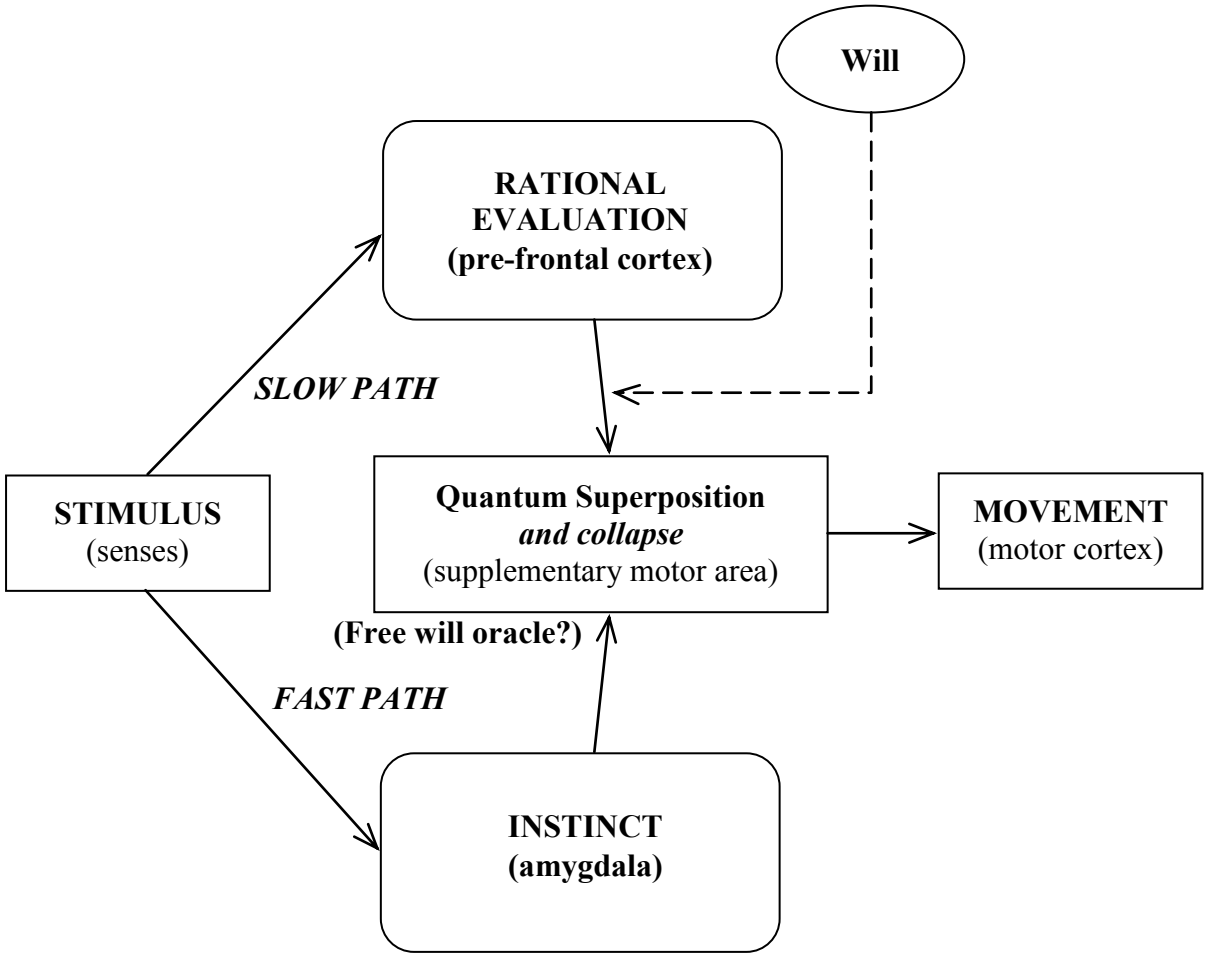


**Figure 1.** Neurobiological flowchart for free choice by an agent. The conflict situation triggers neural signals along two paths: the fast "nature" path and the slow "reason" path-carrying possibly opposing recommendations for action. If the neural signal in the slow path is weak, the agent executes the instinctive action induced by the fast path. But if the signal in the slow path is sufficiently strong, and there is a conflict between $P_0$ and $P_\infty$, then a quantum superposition is set up at the FW oracle, as the intermediate step for deviating $X_0$ towards $X_\infty$.

Some other issues, with ramifications for quantum physics, neuroscience, mathematics, philosophy, computation theory, are briefly mentioned below.

There appears to be a parallelism between the will in $\boldsymbol{F}_1$ and hidden variables in ontological models of quantum mechanics [6]. But there are two basic differences: these ontological models attempt *explain* probabilistic physical laws $W_0$ (and thus correspond to $\boldsymbol{F}_{0,5}$), whereas the will in $\boldsymbol{F}_1$ may produce *deviations* from $W_0$. Second, the will in $\boldsymbol{F}_1$ is the first rung in an infinite hierarchy of higher-order willings, whereas hidden variable models of quantum mechanics stop at unit depth.

Neurobiology is the area most affected by our model, and also its possible clearest testing ground. Experimentally locating the seat of the FW oracle in the brain, and working out how mental causation initiates free-willed action in motor neurons will be vital. Experiments that distinguish willful and casual choice offer another window of study the neurobiological circuitry for FW.

This understanding can be medically useful. By potentially clarifying the roles neurotransmitters or receptors play in the process of free choice, it may be able to suggest

medical solutions that help encourage self-controlled behavior, by enhancing the "reason" pathway, rather than momentarily suppressing the "nature" pathway (Figure 1). Such treatments may be useful for patients suffering from neuropathological ailments like obsessive-compulsive disorder (OCD).

In mathematics and computation theory, the relationship between the causal hierarchy and Turing degrees would merit further study. This will help to elucidate the scope of AI. The formalization of FW as presented here, along the lines of formalization of the concept of truth [3] would be the first step here.

Our model implies that high FW correlates with improved predictability, i.e., reduced entropy in $X_0$. This reduction comes not by a compensatory increase in entropy elsewhere in the universe as required by the Second Law of thermodynamics, but by means of the deviation from $W_0$ produced by mental causation. Departure of $P_j$ from $0_j$ is expressed as a conflict, and thus entropy, on $L_0$. A purely physical means to reduce entropy would be subject to the Second Law, with no implications for the agent's cognitive freedom. But by freely reducing this entropy, the agent is aligning her $P_j$'s with $0_j$'s, and enhancing her freedom. Thus concepts like moral responsibility and justice are helpful as props that encourage free-willed behavior, and thereby help reduce disorder, though not necessarily on the physical level.

This in turn has implications for evolutionary biology. It suggests that the underlying force driving evolution was perhaps not Nature's quest for propagating the species most successful at survival, but instead Nature's quest for greater freedom. Darwin-like incremental evolution evolves quantum matter ($\boldsymbol{F}_{0,5}$ agent) through lower animals lacking a CNS ($\boldsymbol{F}_J$ with finite $J$), and then through higher animals equipped with a CNS but no cerebral cortex (free agents, $\boldsymbol{F}_\infty$), and from there, finally to free-willed agents, $\#\boldsymbol{F}_\infty$. Thus *Homo sapiens sapiens* perhaps already represents the limits of Darwinian biological evolution. The remaining evolutionary journey, towards greater freedom of the will, is now "up to us". It should be accomplished through self-determination.

## REMARK

[1]The Royal Norwegian Academy of Sciences (KNVA), the predecessor of DNVA, posed in 1839 the academic question "Is it possible to demonstrate human free will and self consciousness?"

## REFERENCES


[1] Pink, T.: *Free Will: A Very Short Introduction*. Oxford, 2004, http://dx.doi.org/10.1093/actrade/9780192853585.001.0001,

[2] Davis, M.: *Computability and Unsolvability*. Dover, 1982,

[3] Tarski, A.: *The Semantic Conception of Truth and the Foundation of Semantics*. Philosophy and Phenomenological Research **4**(3), 341-376, 1944, http://dx.doi.org/10.2307/2102968,

[4] Heisenberg, M.: *Is free will an illusion*? Nature **459**, 164-165, 2009, http://dx.doi.org/10.1038/459164a,

[5] James, W.: *The Dilemma of Determinism: The Will to Believe*. Dover, New York, 1956,

[6] Nayakar, C.M. and Radhakrishna, S.: *Gödel, tarski and the conundrum of free will*. In: Balakrishnan, J. and Sreekantan, B.V., eds.: *Nature's Longest Threads*. New Frontiers in the Mathematics and Physics of Information in Biology. World Scientific Publishing Co., 2014,

[7] Schopenhauer, A.: *On the freedom of Human Will*.
1939, translated from German to English in Zucker, M.: *The Philosophy of American History: The Historical Field Theory*. 1945,
[8] Laroudugoitia, J.P.: *Supertask*.
In: Zalta, E.N., ed.: *Stanford Encyclopedia of Philosophy*, 2009,
[9] Pećnjak. D.: *Complex Freedom*.
Interdisciplinary Description of Complex Systems **7**(1), 14-21, 2009,
[10] Lerman, M.: *Degrees of unsolvability*.
Perspectives in Mathematical Logic. Springer-Verlag, Berlin, 1983.