The Journal of Mind and Behavior
Autumn 2022, Volume 43, Number 4
Pages 305–322
ISSN 0271–0137

Free Will, Temporal Asymmetry, and Computational Undecidability

Stuart T. Doyle

Third Force Reconnaissance Company, USMC

One of the central criteria for free will is "Could I have done otherwise?" But because of a temporal asymmetry in human choice, the question makes no sense. The question is backward-looking, while human choices are forward-looking. At the time when any choice is actually made, there is as of yet no action to do otherwise. Expectation is the only thing to contradict (do other than). So the ability to do something not expected by the ultimate expecter, Laplace's demon, is a better criterion for free will. If human action is fundamentally unpredictable, then we have free will. Scientists have studied a form of fundamental unpredictability, known as undecidability. The features that make a system capable of undecidable dynamics have been identified: program-data duality; potential to access an infinite computational medium; and the ability to implement negation. Humans have all three of these features, so we very likely are fundamentally unpredictable, so we have free will.

Keywords: undecidability, predictability, temporal asymmetry

The question of whether or not humans have free will has fascinated many writers for many years. Those denying free will have most often focussed on the supposed incompatibility between free will and deterministic natural laws. These efforts have converged on two main types of arguments:

1. Arguments for the claim that determinism would make it impossible for us to be the source of our actions in the right kind of way.
2. Arguments for the claim that determinism would rule out our ability to do or choose otherwise. (Vihvelin, 2018)

In a recent article, I (Doyle, 2021) critique the first type of anti-free will argument, using the concepts of scale and emergence, which are similar to themes found in List's (2014, 2019) case for free will, which is similar to Ismael's (2013, 2016) case for free will. The idea is that at the scale of the whole human agent,

Correspondence concerning this article should be addressed to Stuart T. Doyle, 3rd Force Recon Co, 1630 S Broad Street, Mobile, Alabama 36605. Email: stuartdoyle1@gmail.com

will exists. At that scale, the will of the human agent is causally relevant and brain molecules are not causally relevant (Doyle, 2021). This paper will build on this concept of scale. I will extend and augment it to refute the second type of argument against free will.

This is the basic form of the argument I will refute in this paper:

a. If someone acts of his own free will, then he could have done otherwise.
b. If determinism is true, no one can do otherwise than one actually does.
c. Therefore, if determinism is true, no one acts of his own free will.

There is already a standard way that free will compatibilists respond to this argument: they make it conditional. They say that a person could do otherwise if he wanted to do otherwise, or if he had a reason to do otherwise (Ayer, 1954; Bok, 1998; Dennett, 1984; Fischer, 1994; Frankfurt, 1971; Hobbes, 1654/1999, p. 16; Hume, 1748/1975, VIII.1; Locke, 1690/1975, II.xx.8; Moore, 1912). This classical version of compatibilism is a well-worn response with an obvious counter: "But it wasn't within your power to want to do otherwise, as what you want is also determined by natural laws." In this paper, I will form a completely different kind of compatibilist response, using an analysis of temporal asymmetry, and applying a concept from computation theory — undecidability — in combination with my description of the agent scale.

Temporal Asymmetry

There is a temporal asymmetry in the question of whether I could have done otherwise. In the question's typical form, it is backward-looking. It asks about what could have been in the past, and at first it seems like a coherent question. I did one thing yesterday, and we wonder if I could have done something else. But what if we wanted to figure out whether or not I'll have free will tomorrow? From that temporal angle, the question of the ability to do otherwise stops making sense. In a forward-looking sense, the question becomes manifestly nonsensical. Can I do otherwise in the future? "Otherwise?" Other than what? Other than the thing I will do? The question stipulates that I will do a certain thing, and simultaneously asks whether or not I can avoid doing that thing. The stipulation contained within the question makes the answer trivial. No, of course I can not do something other than the thing I will do. In order for the question to have any significance in the forward-looking tense, it must be modified. The question can not directly stipulate that I will do a certain thing. The question must ask whether or not I can do something other than what I'm expected to do, not other than what I will do.

In order for the modified question to be significant, "expected" must be taken to mean something stronger than the word's usual usage. If the ability to do other than expected is supposed to work as a criterion for free will, then "expected" here

must mean something stronger than merely what a human might expect that I would do, because surprising a basic human is easy and does not tell us much about free will.

In order to get to the bottom of a forward-looking notion of free will, we need a non-trivial form of the question: Can I do otherwise? And for that, we need a significant "what," as in "other than what?" For that purpose, we need the strongest possible notion of what's expected, but not directly stipulated. The strongest conceivable form of expectation would be that of a being who knows the present masses, positions, shapes, temperatures, velocities, charges, spins, and all other physical properties of all particles in the universe; and has the ability to analyze these data. If a being like that expected a certain action to happen, that expectation would really mean something. Such a being was imagined by Leibniz in 1680, and by Laplace in 1820 as illustrations of their views of determinism. The imagined being has since come to be called Laplace's demon. According to Laplace, his demon could perfectly predict the future state of anything in the universe, large or small (Laplace, 1820/1951). Since this paper will not claim indeterminism as a source of free will, let's ignore for the sake of argument the quantum mechanical uncertainty principle, and grant to Laplace's demon the perfect knowledge that's denied to us mortals who live under Heisenberg's regime of uncertainty. The demon is the ultimate predictor. What he expects is the ultimate expectation. What Laplace's demon expects me to do in the future is the "what" in "other than what?" So the forward-looking "Can I do otherwise?" is "Can I surprise Laplace's demon?" If I can surprise the demon, then I have free will by the forward-looking criterion of the ability to do otherwise.

But this raises another question: In judging a choice to be free or unfree, should we use the forward-looking formulation or the backward-looking formulation of the ability to do otherwise? We should use only the forward-looking formulation because a choice by its nature is forward-looking. We don't deliberate or make choices about the past. For details and implications of this fact see Fernandes (2017). A choice is an event which coincides with the end of deliberation. Ensuing events are determined by the deliberative choice process. We only make choices about the future, not about the past. Choices at the human scale do not have time reversal symmetry. Though there are many physical processes at smaller scales that do have time reversal symmetry, that is not relevant to human choice. More will be said about scale and time reversal symmetry in the next section. For now the point is that in human choice, the choice is always *about* something, and that object of choice always lies in the future, thus choice is always forward-looking.

At the time when a choice is actually made, there is as of yet no "what" as in "Could have done other than what?" I have not already made the choice, so there is no established action to have done otherwise. But there is expectation. When I make a choice, there are options which seem open to me. There can be expectations of which option I will choose. The expectation is the only thing which I

might contradict (do otherwise) at the time of my choice. So in analyzing the actual event of a choice, the question about the ability to do otherwise must be forward-looking. For the question to make any sense, it must be "Can I do other than expected by the ultimate expecter, Laplace's demon?"

One might object that we can still pose coherent backward-looking questions about forward-looking choices. We can talk about a choice I made yesterday even though the choice itself was inherently future oriented. But there are some questions we actually can not ask in this way. Relevant to this discussion, we can not ask what it's like to exist and move through this world. As far as we know, what it's like will always have the quality of moving forward through time. The question of free will is supposed to be about the way in which a human exists and acts in this world. The question is, "When I make a choice, is it free? The only way to properly answer this question is to talk about the choice as it is *when I make it*. Every instance of choice occurs at a time when there is as if yet no "what," as in "other than what?" So we should use only the forward-looking formulation of the ability to do otherwise.

So far, I've considered the forward-looking and backward-looking forms of the ability to do otherwise. There is also a formulation which seems to situate the ability in the present. It seems that one could ask, "Is there more than one action I am now able to perform?" Van Inwagen (1983, p. 8) framed the ability to do otherwise in such a way. But this formulation makes no sense regardless of determinism. If this formulation is meant to stay in the present, and not collapse into either the forward-looking or backward-looking formulation, then the alternative actions must be in the present. But any action of mine which is actually in the present is what I am doing right now. Of mutually exclusive actions (i.e., speaking or being silent), I can only be presently performing one. So the "actions" in "Is there more than one action I am now able to perform?" must not be present and actual, they must be future and speculative, or else I would be doing them — but not all of them, just one of them. There can be no such thing as multiple possibilities which are truly in the present, since we *are doing* whatever is possible in the present. So any talk of multiple possibilities is referring to the future, not the present. And if these multiple possibilities are the objects of a choice, they may only exist in the future which follows the choice, not in the present at the time of the choice.

For speaking and being silent to both be possibilities, they must be in the immediate future, not literally "now." So before considering determinism, we should amend van Inwagen's question from "Is there more than one action I am now able to perform?" to "Is there more than one future action which I might perform?" Speaking and being silent in the near future are both epistemic possibilities in that we don't know which will happen. They are both logical possibilities, in that neither action would be logically self-contradictory. So the question is, are they both possible under the actual laws of nature?

In the forward-looking sense, a future action is possible under the actual laws of nature if Laplace's demon can not rule it out before it happens or fails to happen. The only other way to judge physical possibility would involve a backward-looking analysis; i.e., "at a certain time in the past, I spoke. And since my actions followed from natural laws, it must have been impossible for me to be silent at that time." As argued above, the nature of deliberative choice is incompatible with backward-looking analysis, so the actions which follow from a deliberative choice should only be judged to be possible or impossible by the forward-looking criterion.

So when temporal asymmetries are taken into account, the ability to do otherwise and the ability to "now" perform more than one action can both be formulated in terms of Laplace's demon. The ability to do otherwise is the ability to do something that the demon does not expect. The ability to "now" perform more than one action means that the demon can not rule out all but one action in my immediate future. Since the demon is always right, these two are equivalent. If I can do something that the demon does not expect, that means the demon does not have a firmly established expectation of what I will do, which means that the demon can not rule out all but one action.

With temporal asymmetry taken into account, we have a refined criterion for free will: "Can I choose to do something not expected by the ultimate expecter, Laplace's demon?" The answer to this question will turn out to be yes. I can surprise Laplace's demon, and thus I can do "otherwise," and thus I am free. But there are a few more steps of reasoning required to reach this conclusion. It starts with a consideration of the human agent at the proper scale. This notion of human scale was recently developed in my 2021 article. Similar ideas are found in List (2014, 2019) and Ismael (2013, 2016), but as mentioned above, primarily my version will be described here as a necessary background for answering the question of demon surprise. Both List and Ismael have fleshed out their versions in ways that do not fit neatly with the direction of this paper.

Scale

Many of our actions are caused by our wills; that is, by our conscious desires and intentions. This is usually not disputed by most free will deniers. They more often dispute that our wills are free, not that we have wills and that our actions often follow from our wills. Sam Harris (the free will denier most popular with the general audience) has said that the subjectively felt intention to act is the proximate cause of acting. Harris makes the same basic claim made by Crick (1995, p. 3), philosophers such as Pereboom (2001, p. 112), and many before them. They claim that in addition to the proximate cause (the will), our actions have more ultimate causes which are the relevant causes to consider when judging whether or not our wills are free. The ultimate causes beyond and beneath the surface of

our wills supposedly make them unfree. So what are these ultimate causes? Harris recently identified genetics and environmental influences as "the only things that contrive to produce" his particular will (Harris, 2021, 69:00). While Harris finds DNA particularly relevant, similar reductive views submit other types of molecules as the hidden authors of our choices. For example, Jerry Coyne voices the intuitive critique of free will made by many college freshmen: "Our brains are made of molecules; those molecules must obey the laws of physics; our decisions derive from brain activity" (2019, ¶ 4).

So what's wrong with this line of thinking which is so drawn to molecules and such? Consider the following question as an analogy: Are apples red? Suppose we all agree that apples have color. The question is whether the color is red or non-red. To answer the question, Sam Harris and Jerry Coyne look beyond the proximate color of the apple. Realizing that the apple is nothing but atoms, they examine many of the carbon atoms on the surface of the apple. They find that not a single carbon atom is red. Since none of the atoms are red, and the apple is nothing but atoms, Harris and Coyne conclude that the apple can't be red. Their error is that though they agree the apple has a color, they try to examine the nature of the color at a scale where color is incoherent. (A carbon atom is smaller than the wavelength of red light.) The fact that they found no redness at that scale shouldn't lead them to conclude anything about the redness of the apple.

Likewise, the fact that Harris and Coyne find no personal authorship or freedom in the actions of molecules shouldn't lead them to conclude anything about the nature of the will. We agree that we have wills, that we have subjectively experienced intentions which cause our actions. The question is whether the will is free or unfree. To look at molecules for the answer is a mistake. DNA and neurotransmitters observed at the molecular scale exhibit no will whatsoever. With that knowledge, should we really find it compelling that molecules exhibit no *free* will? No. It should tell us that Harris and Coyne are looking at the wrong scale to find answers about the will, just like looking for answers about redness at a scale where there is no color.

The right scale for finding answers to the question of apple redness is the apple scale, not the atom scale. The right scale for finding answers to the question of freedom of the will is the agent scale, not the molecule scale. Searching the molecule scale is just one example of this error. There are many other wrong scales where a confused free will denier might erroneously search for answers about the will. He may zoom out temporally into an irrelevant timescale, including the time before the will in question existed. In the analogy, this would be like conceptualizing the apple as merely a step in a process of agricultural industry. Since agricultural industry is not red, should we conclude that the apple is not red? No, we should realize that the question about the apple should only find its answers from a scale where the apple exists as an apple. And the question about the will should only find its answers from a scale where the will exists as a will. To declare

a lack of freedom by appealing to the time before the agent was born is to lift descriptors for the will from a timescale where no will can exist. It is as incoherent and irrelevant as pointing out that the agricultural industry is not red.

If we keep our analysis in the scale where the individual agent exists, not zooming too far in nor too far out in space, time, or level of organization, then the primary and ultimate cause of my actions is me. The will emerges from the complex interactions of many small parts. It's literally not true to say that it's caused by any particular small part. It is caused by many small parts, but *only* when taken together all at once. And that's the same thing as the whole person. So my thoughts and actions are deterministically caused by me. The molecules of which my brain is made are deeply irrelevant to this fact.

Besides obscuring the true source of human actions, an inappropriate focus on the dynamics of little particles could also obscure the truth of temporal asymmetry in human choice. The laws of physics that describe or govern the interactions of particles do not specify a direction of time. If we could watch a video of two protons colliding, we would have no way to know whether the video was being played forward or in reverse. This is called time reversal symmetry. This symmetry holds true in a wide variety of particle interactions (Carbone and Rondoni, 2020). Time appears asymmetric only at scales where emergent phenomena transpire. Large collections of particles obey the second law of thermodynamics, which is not time reversal invariant. As astrophysicist Matt O'Dowd puts it, "Zoom in to individual particle interactions and you see the perfect reversibility of the laws of physics. But zoom out, and time's arrow emerges" (O'Dowd, 2020, 6:32). So it turns out that the temporal asymmetry described in the previous section depends on the understanding of scale described in this section.

Because of the different dynamics found at different scales, the ability to do otherwise needs to be understood as temporally asymmetric; that is, as always forward-looking; as the ability to do something not expected by Laplace's demon. The path to understanding this ability is also illuminated by consideration of scale, which leads to an understanding of the self as the true source of one's actions.

Self-Reference

The fact that I am the relevant cause of my own actions comes with another important implication: I am a causally self-referencing entity. If a molecule were the relevant cause of my action, this would not be true in the same way. The molecule has no capacity for self-reflection, but I do. I can ask myself, "What will I do? What could I do? What should I do? What do I want to do? What should I want to do? What would I do if I wanted to do x and should do y? What would I become if I did x? What would I do if I became that thing that results from doing x?" Self-referential questions like these affect the choices that I make; and those choices change the self-referential questions that I ask.

Undecidability

At the relevant scale, self-reference is causally important. I am a system which analyzes its own inputs, character, and potential outputs; generates new outputs based on those analyses; and feeds those new outputs back into itself as inputs which affect the outputs, which affect the system's character. I am an output of and an input for my own processing. Framing the human self-referential nature in this way brings us to the next step in surprising Laplace's demon: computational undecidability. This is a term which describes a system which can not be predicted, given complete knowledge of its present state. This fundamental unpredictability shows up in algorithmic computation, formal mathematical systems, and dynamical systems. Though an unpredictable dynamical system may evoke the concept of chaos, undecidability is a different sort of unpredictability.

> For a dynamical system to be chaotic means that it exponentially amplifies ignorance of its initial condition; for it to be undecidable means that essential aspects of its long-term behaviour — such as whether a trajectory ever enters a certain region — though determined, are unpredictable even from total knowledge of the initial condition. (Bennett, 1990, p. 606)

If a system exhibits undecidability, then it is unpredictable even to Laplace's demon, while a system that is merely chaotic is perfectly predictable to the demon. So what's left for me to do in my project of describing the human ability to do otherwise is to make a case that we humans are systems which exhibit undecidability. To do this, I'll apply three criteria that complexity scientists identify as characterizing the underlying logic that generates undecidability. In 2019, Mikhail Prokopenko and coauthors conducted a comparative formal analysis of recursive formal (mathematical) systems, Turing machines, and cellular automata. They come to a clear conclusion: "As we have shown, the capacity to generate undecidable dynamics is based upon three underlying factors: (i) the program–data duality; (ii) the potential to access an infinite computational medium; and (iii) the ability to implement negation" (p. 154). Now I'll describe in turn what program–data duality, infinite computational medium, and negation are; and why humans should be thought of as having these properties. If humans do have these three properties, then we meet the criteria for undecidable dynamics, which means we can take actions not expected by Laplace's demon, which means we have free will.

Program–Data Duality

Program–data duality in this context is the ability for self-reference (Prokopenko et al., 2019, p. 143). The word "duality" refers to the typical distinction between program and data with which we are all familiar. For a simple example, a

pocket calculator has a program: its set of rules. It does not yet know which buttons will be pressed; those are the data. A human at time t1 has a certain overall state of mind, coinciding with a certain overall physical state. The state at t1 is a program, in that it entails implicit rules about what the system would do, given certain types of data. The streams of perceptions taken in at t2 are data, which get processed according to the implicit rules. In addition to processing basic sense data, this duality allows for a program (or implicit set of rules encoded in the state of a human) to process other programs as data. For example, a human can process ideas, hypothetical scenarios, mathematical operations, and representations of the self as data. As the complexity researchers put it, "Undecidability arises due to the self-referential ability [of a system] to interpret and run an input which encodes its own description, reflecting the program–data duality" (Prokopenko et al., 2019, p. 143). Since self-reference is causally important in humans, we meet this requirement.

Infinite Computational Medium

The next requirement for undecidability is the potential to access an infinite computational medium. The computational medium is the substrate on which the state of the system is represented. In a Turing machine, this is the tape. In a cellular automaton, this is the grid lattice on which cells may be white or black. The set of all possible states of the system is called the state-space. An infinite computational medium accommodates an infinite state-space. If the computational medium is finite, then the state-space must be finite. For example, a cellular automaton with a 2 x 2 grid lattice has a state-space of size $2^4 = 16$, meaning that there are only 16 distinct ways in which the grid may be tiled in black and white. So if we knew that a certain system had an infinite state-space, we could infer that the system has access to an infinite computational medium.

It can be informally shown that humans have a state-space of infinite size. Any natural number can be thought of by a human. Even the large numbers that have no obvious relationship to everyday life can be conceived of in their relationships to other numbers. Each conception of a number must be a different mental and physical state than the conception of any other number. One may doubt that literally any natural number may be thought of. What about a number with a thousand digits and no repeating patterns? Such a number can not be held in the mind. But such a number could be read off to a human, and the reading of the number would result in a certain impression, despite the human not holding most of the digits of the number in her mind at any one time. If a different number were to be read, a slightly different impression would result. The point here is not that a person is able to form impressions of each and every number; that would not be possible. The point is that a person can form an impression of any single number. And such an impression would be unique to that number. If there are

infinite numbers, then an infinite number of unique impressions are within the state-space of the human.

Referring only to numbers makes the example simple, but of course the state-space of the human is far larger than the space of conceivable numbers. Think of the number 74. Now think of the number 74 with your eyes closed. Those two occasions of thinking of 74 occupied two very different points in your state-space because of the difference in visual perception. How many different visual fields might a person be able to perceive while thinking of 74? To roughly estimate how many states are possible while thinking of 74, we would need to do something like multiply the number of possible visual perceptions by the number of possible auditory perceptions by the number of possible olfactory perceptions by the number of possible sensations of heat and cold by the number of possible gradations of feeling sadness or happiness, and so on. Also, you may think of 74 while remembering the time you thought of 106 or 107, and so on. And the next time you think of 74, that will be yet another point in your state-space, since you'll recall that you've thought of 74 before. There may be an infinite number of states associated with thinking of 74. And there are many conceivable numbers other than 74, and many things to think about other than numbers.

An obvious objection might be that a human and his brain are physically finite. In what sense can an organ that fits inside a skull be infinite? As a starting point, consider the 100 billion neurons that make up the brain. As a simplification, a neuron can be considered to be "firing" or "not firing." So a simplified brain has 100 billion binary cells. Such an array of cells could instantiate $2^{100,000,000,000}$ distinct patterns of on-or-off activation. That's a big number. For reference, there are estimated to be roughly 10^{80} atoms in the observable universe (Padilla, 2017). The number of atoms in the universe is an infinitesimally small number compared to the number of activation patterns possible in a simplified brain. And what about a real brain? A real brain is made of neurons which are not simply on or off. Some neurons show analog gradations in voltage and neurotransmitter release, meaning that they have many possible states between "on" and "off" (Zbili et al., 2016). Besides analog neurons, there are many variables in the brain which are also not captured by the simplified on/off digital variable. Each neuron can vary in the amount of neurotransmitter in its vesicles ready for release, and the state of the receptors on its soma and dendrites (to what degree they're blocked by other molecules). There can also be variation in the amount of neurotransmitter which is floating free at any moment in the space between any two neurons. There are also minute variables which will likely never be measured, yet do theoretically make a causal difference. For example, in what spatial direction is each neurotransmitter molecule oriented? A neurotransmitter molecule must fit into a receptor in order to carry on a signal. For the molecule to fit, it must be facing a certain direction relative to the receptor. So the spatial orientation of the molecule before binding must have some nonzero effect on the binding affinity. How many

different patterns of analog spatial orientation might trillions of neurotransmitter molecules be capable of? This alone may be infinite. The digital variable of "firing" or "not firing" does not capture any of these variables.

With the complex interaction of the digital and analog factors I've mentioned, as well as many factors not mentioned, the medium of the brain may actually be computationally infinite although it is finite in mass and volume. Whether the human state-space is technically infinite or merely practically infinite (larger than any other number computed for any purpose in all of science), it will not be exhausted in the meager 100 years of a human lifespan. This means that the self-referential loops of processing do not need to stop at any predetermined iteration or level of abstraction. So for the purpose of analyzing the choices of a human, the state-space and computational medium are functionally infinite.

Negation

The last element required for undecidability is the ability to implement negation. Negation in this context refers to the ability of a logical system to produce an output which is exactly contrary to the processing which led to the output. It is equivalent to the liar paradox, which is exemplified in a statement such as "everything I say is a lie," or more formally, "this statement is unprovable." The liar paradox is a self-referential statement, which can not be judged to be true or false without a contradiction. Self-reference is fundamental to this paradox because the statement refers to its own validity. If humans can implement this paradoxical logic into their thinking, then humans meet this requirement for producing undecidability. The fact that humans came up with the liar paradox thousands of years ago is evidence that humans can perform the logical operation of negation.

Conclusion

So all three factors underlying the capacity to generate undecidable dynamics are present in humans. Humans exhibit program-data duality when we process ideas, hypothetical scenarios, mathematical operations, and representations of ourselves as objects of thought. We have the potential to access an infinite computational medium. This is demonstrated by the fact that we can think of any one of an infinite number of objects of thought, which implies an infinite state-space, which implies an infinite computational medium. We have the ability to implement negation, demonstrated by the inception of the liar paradox in the minds of humans. If these three elements are sufficient to generate undecidable dynamics, then humans are capable of generating undecidable dynamics, which means we can not be accurately predicted by Laplace's demon. And that means we have the ability to do otherwise in the forward-looking sense, which is the only formulation that works as a coherent criterion for free will.

Figure 1 shows the relationships between the concepts discussed in this paper. An understanding of the human agent at the scale where conscious humans actually exist leads to recognition of the self as the source of one's actions, recognition of the relevance of temporal asymmetry to human choice, and recognition of self-reference as causally relevant to human actions. Self-reference, also called program-data duality, in combination with access to an infinite computational medium and the ability to implement negation, results in undecidable dynamics. That is, fundamental unpredictability. That unpredictability entails the ability to do otherwise in the forward-looking sense, which is the only sense that makes any sense when temporal asymmetry is taken into account. The resultant total picture is that of two criteria for real free will which are met by human agents: the forward-looking ability to do otherwise, plus being the source of one's own actions.

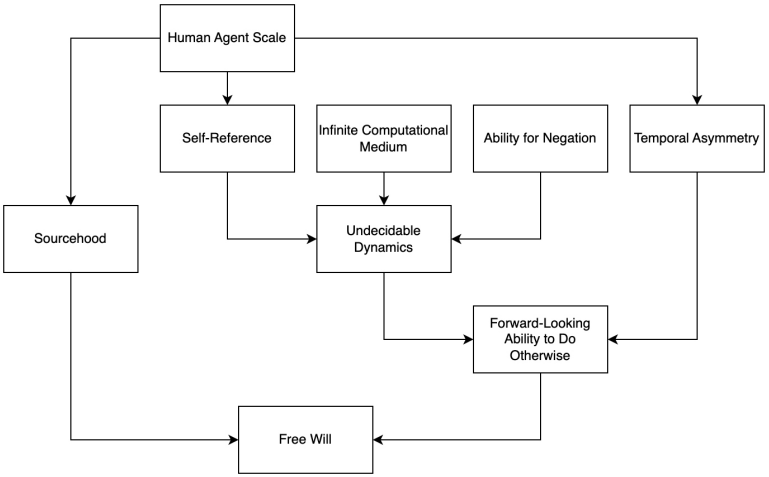

Figure 1: Relational map of concepts. The truth of each concept supports the truth of the concepts downstream from it. This diagram illustrates how the concepts described throughout this paper contribute to the overall view developed here.

Each step in the project of this paper relies on informal argumentation. I can not formally (mathematically) prove that humans exhibit undecidability. In order for anyone to do that, there would need to be some known algorithm which accurately represents a human. No such algorithm is remotely close to being known. It would be absurdly complex. Even the underlying algorithms of drastically simplified analogs such as roundworms and simulated neural networks are elusive to researchers (Lynn and Bassett, 2019, p. 324). Formal proofs of undecidability in physical systems far simpler than a human have been done, and they are monumental works of mathematics. I'm referring to Toby Cubitt, David Pérez–García, and Michael Wolf's (2015) proof of the undecidability of the spectral gap.

The unknown dynamics of the human system make such a formal proof impossible. But it is feasible to identify in humans each of the essential factors responsible for undecidability. If these factors are sufficient for undecidability, as shown by Prokopenko et al., then human choices should be expected to exhibit undecidability.

The connection between predictability, determinism, and free will has previously not been clear. The original point in Leibniz's and in Laplace's ideas of perfect predictions was that determinism implies predictability. Such predictability seems to make people feel a threat to free will (Hoefer, 2016, § 1). Alison Fernandes has suggested that ignorance of our own future actions is what makes us feel free, and thus that predictability would eliminate at least the feeling of freedom (Fernandes, 2019). Harris has taken a Laplacian notion of predictability to evoke a lack of free will, though most free will deniers including Harris don't directly use predictability in their arguments (Harris, 2021). It seems that the idea of predictability gets people to ponder the question of free will, and the assumption of either predictability or unpredictability leads to a feeling of freedom or a feeling of unfreedom respectively. As predictability has seemed to weigh against free will, though not decisively, unpredictability has seemed to weigh in favor of free will, though not decisively. Hilary Bok (2007, p. 138) refers to unpredictability as epistemic freedom, which she defines as something different from metaphysical freedom, which is actually being able to do otherwise. She argues that the epistemic sense of freedom is of far more practical importance than the metaphysical sense, and so we should consider ourselves free, though not metaphysically. In contrast to Bok, what I am doing is collapsing unpredictability and metaphysical freedom into one. Because of the temporal asymmetry in the natures of deliberation, choice, and the ability to do otherwise, the strongest form of epistemic freedom (the ability to surprise Laplace's demon) is the only thing that can be meant by "the ability to do otherwise." So strong unpredictability does imply metaphysical freedom, but not indeterminism.

There are views in which determinism and predictability are both said to be eliminated in the context of human choice by quantum indeterminacy. But critics of these views point out that if the relevant cause of an action is an indeterminate quantum event, then the human agent can not determine what he does, and thus can not be the source of his own actions (Pereboom, 2014, p. 32). I agree with the critics on this point. In contrast to quantum indeterminacy, undecidable dynamics are deterministic, and are a property of the human system taken as a whole, not a property of some little part of a human. So undecidability fits with the notion of humans determining their own actions. Thus one idea, the scale dependent view, leads to mutually congruent answers to both central questions in the free will debate: sourcehood and the ability to do otherwise. Viewing human agents as whole humans instead of as molecules makes it clear that the human agent is the cause of her own actions, and also leads to a focus on the human features such as self-reference, which underlie undecidable dynamics.

Since I endorsed the classical compatibilist line[1] in a previous article (Doyle, 2021, p. 286), it may seem that adding another answer (undecidability) signals a lack of confidence in either answer. But this is not quite true because both answers follow from one principle: self-reference. Properly considering the human scale where will exists leads to the conclusion that I am the source of my own actions. That is, that self-reference is the correct way to understand the causality of one's choices. The classic conditional answer to the question of the ability to do otherwise is simply a direct application of self-referential causality to the question as stated. It turns out, as revealed in this paper, that the question is typically stated in a confused backward-looking way. This may be the reason why the same classical response has been made, rebutted, and remade for centuries without either side conceding defeat. A malformed question may never be satisfactorily answered. When the question of the ability to do otherwise is reformulated to be temporally coherent, the principle of self-referential causality can still be applied. But now it leads to undecidability. This provides a conclusion with more finality than the classical compatibilist response. The trajectory of human thought is either computationally undecidable, or it is not. If it is, then we do have the ability to do otherwise, and so we do have free wills.

The view put forth in this paper also has advantages in that it is not vulnerable to strong critiques which apply to other scale-based accounts of free will. John Daniel Wright (2022) recently published a refutation of List's scale-based libertarian account of free will. According to List, the behavior of human agents is undetermined, while the behavior of little bits of matter is determined. Wright rightly points out that if the atoms in my body are determined by physical laws to be in London on Saturday, then I can not nondeterministically choose to be in Manchester, even though I feel like I can choose to go there on Saturday. This example makes very clear the correspondence between the fate determined for the atoms and my supposedly undetermined choice. Can an undetermined outcome have such clear correspondence with determined outcomes? Probably not. But this is not a problem for the undecidability based view of free will.

An atom in my body is determined to be where it is by interactions with the atoms adjacent to it. And those atoms are determined to be where they are by interactions with the atoms adjacent to them, including the first atom I already mentioned. If we want to figure out where all those atoms will be on Saturday, we need to consider the interactions between all of the atoms in my body, including those which make up my nervous system. But the behavior of that system of atoms is undecidable. Its trajectory can not be specified ahead of time, so the trajectory of the single atom can not be specified either. And so there is no conflict at all between the atom's determined behavior and my agential choice, which was

[1] "I could have done otherwise if I wanted to."

determined by me and unspecifiable before it came to be. The atom's trajectory was also unspecifiable in the same timeframe.

Wright also gives a second example that shows the problem with List's account:

> Imagine we have a small physically deterministic system where atoms fired at a slit can either pass through the slit or instead impact the surface around the slit. If the atom passes through, the machine, at the macro level, turns on a red light that scientists watching the process can observe and so they know the atom passed through. If, after the atom is released, it hits the surface, a green light comes on instead. Although the system is physically deterministic, for reasons of calibration and other micro-variables, sometimes the atom passes through and sometimes it does not. In List's model, we have an indeterministic system at the macro level, with both red light events and green light events being equally possible. (Wright, 2022, § 3.2)

To be consistent with List's own views, List would have to say that when the red light comes on, the green light could have come on. But the atom could not have taken a different path, and the path taken by the atom is what decides whether the lights come on red or green. If Wright has rightly characterized List's view — and I think he has — then List's view leads to contradictions, revealed by these examples. But no problem is posed here for my view. Quite simply, the atomic slit apparatus does not generate undecidable dynamics, but humans do. Wright's apparatus could be predicted with perfect accuracy by Laplace's demon. There is a principled difference between Wright's apparatus and a human agent: humans can surprise Laplace's demon.

To conclude, I offer a closing summary of the points I've made in this article. Humans make decisions and act. The little mechanical parts that make up a human body are not the relevant sources of those decisions. The emergent whole person is the source. Events and causes at the human scale, unlike events at very small scales, are temporally asymmetric. Importantly, human choices always happen before the actions which are chosen. Choices are always *about* something, and that object of choice is always in the future. One of the central issues of free will, the ability to do otherwise, is all about choice. Since choices are always forward-looking, the ability to do otherwise must be forward-looking. That is, the question must not be about the past, but about the future. And a question about potential actions in the future must not as a stipulation hold fixed some action in the future. Otherwise, there is no coherent question at all. The question must not posit that a person absolutely *will* do a certain thing in the future, and ask whether she can do otherwise. The question must ask whether a person can do other than *expected*. "Expectation" can be taken to the extreme sense of Laplace's demon, in order to find the in-principle answer to the question of free will. If we can surprise Laplace's demon, then we should say that we have the ability to choose otherwise. We can surprise Laplace's demon because at the scale of the human agent, we have

the three intrinsic characteristics which produce undecidable dynamics: self-reference, infinite computational medium, and the capacity for negation. Since the trajectories of our thoughts and actions are undecidable, we are fundamentally unpredictable, and so we do have the ability to do otherwise.

References

Ayer, A. J. (1954). Freedom and necessity. In *Philosophical essays* (pp. 3–20). New York: St. Martin's Press.

Bennett, C. H. (1990). Undecidable dynamics. *Nature, 346*, 606–607.

Bok, H. (1998). *Freedom and responsibility*. Princeton University Press.

Bok, H. (2003). Freedom and practical reason. In G. Watson (Ed.), *Free will* (pp. 130–166, second edition). Oxford University Press.

Carbone, D., and Rondoni, L. (2020). Necessary and sufficient conditions for time reversal symmetry in presence of magnetic fields. *Symmetry, 12*(8), 1336.

Coyne, J. (2019, July 17). Why we shouldn't bet on having free will — a reply to William Edwards. *Quillette*. Retrieved from https://quillette.com/2019/07/17/why-we-shouldnt-bet-on-having-free-will-a-reply-to-william-edwards/

Crick, F. (1995). *Astonishing hypothesis: The scientific search for the soul*. Scribner.

Cubitt, T. S., Perez–Garcia, D., and Wolf, M. M. (2015). Undecidability of the spectral gap. *Nature, 528*, 207–211.

Dennett, D. (1984). *Elbow room: The varieties of free will worth wanting*. MIT Press.

Doyle, S. T. (2021). Sizing up free will: The scale of compatibilism. *Journal of Mind and Behavior, 42*(3 & 4), 271–290.

Fernandes, A. (2017). A deliberative approach to causation. *Philosophy and Phenomenological Research, 95*(3), 686–708.

Fernandes, A. (2019, February 22). The future seems wide open with possibilities — but is it? *Aeon*. Retrieved from https://aeon.co/ideas/the-future-seems-wide-open-with-possibilities-but-is-it

Fischer, J. M. (1994). *The metaphysics of free will: An essay on control*. Blackwell.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, *68*(1), 5–20.

Harris, S. (2021). Sam Harris: Consciousness, free will, psychedelics, AI, UFOs, and meaning. *Lex Fridman podcast* #185 https://www.youtube.com/watch?v=4dC_nRYIDZU&t=0s

Hobbes, T. (1999). Of liberty and necessity. In V. Chappell (Ed.), *Hobbes and Bramhall on liberty and necessity*. Cambridge University Press. (originally published 1654)

Hoefer, C. (2016). Causal determinism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/spr2016/entries/determinism-causal/

Hume, D. (1975). *Enquiries concerning human understanding and concerning the principles of morals* (third edition; P.H. Nidditch, Ed.). Oxford: Oxford University Press. (originally published 1748)

Ismael, J. (2013). Causation, free will, and naturalism. In H. Kincaid, J. Ladyman, and D. Ross (Eds.), *Scientific metaphysics* (pp. 208–235). Oxford: Oxford University Press.

Ismael, J. (2016). *How physics makes us free*. Oxford: Oxford University Press.

Laplace, P. (1951). Essai philosophique sur les probabilités. In *Théorie Analytique des Probabilités*. F.W. Truscott and F.L. Emory (Trans.). New York: Dover. (originally published 1820)

Leibniz, G. (1956). *The Leibniz–Clarke correspondence (*H. G. Alexander, Ed.). Barnes and Noble. (originally published 1680)

List, C. (2014). Free will, determinism, and the possibility of doing otherwise. *Noûs, 48*(1), 156–178.

List, C. (2019). *Why free will is real*. Cambridge: Harvard University Press.

Locke, J. (1975). *An essay concerning the human understanding* (P. H. Nidditch, Ed.). Oxford University Press. (originally published 1690)

Lynn, C., and Bassett, D. (2019). The physics of brain network structure, function and control. *Nature Reviews Physics, 1*, 318–332.

O'Dowd, M. (2020). The arrow of time and how to reverse it. *PBS Space Time*, Nov 18. Retrieved from https://www.youtube.com/watch?v=QkWT-xMTm1M

Moore, G. E. (1912). *Ethics*. Clarendon Press.

Padilla, T. (2017). How many particles in the universe? *Numberphile*. Retrieved from https://www.numberphile.com/videos/how-many-particles-in-the-universe

Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford: Oxford University Press.

Prokopenko, M., Harré, M., Lizier, J., Boschetti, F., Peppas, P., and Kauffman, S. (2019). Self-referential basis of undecidable dynamics: From the liar paradox and the halting problem to the edge of chaos. *Physics of Life Reviews, 31*, 134–156.

van Inwagen, P. (1983). *An essay on free will*. Clarendon Press.

Vihvelin, K. (2018). Arguments for incompatibilism. In *The Stanford encyclopedia of philosophy* (E. N. Zalta, Ed.). Retrieved from https://plato.stanford.edu/archives/fall2018/entries/incompatibilism-arguments/

Wright, J. D. (2022). Compatibilist libertarianism: Why it talks past the traditional free will problem and determinism is still a worry. *Journal of the American Philosophical Association, 8*(4), 604–622.

Zbili, M., Rama, S., and Debanne, D. (2016). Dynamic control of neurotransmitter release by presynaptic potential. *Frontiers in Cellular Neuroscience, 10*. Retrieved from https://www.frontiersin.org/articles/10.3389/fncel.2016.00278/full