Is Libertarian Free Will an Inescapably Incoherent Concept?

SAMUEL MICHAELIDES

Thesis submitted to the University of Hertfordshire in partial fulfilment of the requirements of the degree of Doctor of Philosophy (PhD)

February 2024

Acknowledgements

This has been a long road and no doubt there are a number of people who deserve thanks for their help and support. To my parents, Carole and Andrew Michaelides, for your continued patronage (and for so many other things) I will forever be grateful. To my supervisors, Craig Bourne and Sam Coleman, your guidance and generosity has been invaluable. To Craig, in particular, without your constant encouragement and gentle suggestions I very much doubt I would ever have made it to the end. To my children, Jake and Jessica, your complete lack of care about what I have been spending much of my spare time on for all these years and your frequent reminders that it is you two that are the most important has always helped to keep things in perspective.

And, finally, to my wife Nina, who really has had to be the one to put up with the highs, lows, and stress of it all, thank you.

For Rachael Browne. You started me on this path many years ago and gave your time, your counsel, and your friendship; never asking anything in return. I'm only sad you could not still be here to see what you have helped me achieve. I know you would have been thrilled.

Abstract

In this thesis I examine whether libertarian theories of freedom and responsibility should be considered to be inescapably incoherent.

Despite arguably having an intuitive appeal, the libertarian approach – with its requirement to reconcile a particular understanding of free will with an indeterministic universe – is often considered to be unintelligible due to its necessary focus on undetermined acts and its attempts to posit a form of non-necessitating control, which its proponents then wish to claim is sufficient for a justifiable ascription of responsibility. Critics claim that libertarian theorists can only ever be led into what I have called the *Catch-22 of Libertarianism*, whereby undetermined acts can only be the result of random processes for which the agent cannot be considered responsible and any attempts to remove this randomness by positing determining factors can only undermine the libertarian position and negatively affect their own stated requirements for ascribing responsibility.

Despite these issues, a number of philosophers remain committed to developing libertarian models for free action and, more recently, researchers working in the various sciences have also joined the debate on the side of libertarianism. I therefore consider six specific libertarian theories in the attempt to answer my main question concerning the coherence of the position. Three of these are from philosophers, who each represent one of the main strands of libertarian thought: Robert Kane's event-causal theory, Carl Ginet's non-causal theory, and Timothy O'Connor's agent-causal theory. The other three are from scientists who each take a different approach to the problem: Peter Ulric Tse's Criterial Causation theory, Roger Penrose and Stuart Hameroff's Orchestrated Objective Reduction theory, and Henry Stapp's Quantum Interactive Dualism theory. In order to properly (and systematically) assess each theory – and give proper consideration to whether each manages to meet all the goals of the libertarian project in a coherent manner – I have developed what I call my *Criteria for Coherence*, which contains six key principles that I believe fully represent the stated aims of all the researchers whose work I consider. It is my contention that any libertarian theory needs to adequately meet all six principles if it is to be considered fully coherent yet, after assessing each theory against the *Criteria*, I ultimately conclude that *none* of them manage to do so.

This done, I move on to consider what it is about the *Criteria* that makes adhering to its principles so problematic and argue that it is in fact the explicit focus on incorporating indeterminism into their models that is the libertarian's undoing. I go on to show that, despite what they might believe, indeterminism is not actually required to meet either of their key conditions for free and responsible action. I conclude by arguing that both indeterminism *and* determinism should be considered irrelevant to the debate on freedom and responsibility and that the focus should instead be on expanding the explanatory network of the concept of agency, irrespective of any questions over the fundamental physics of the universe.

Table of Contents

Introduction

Libertarian theories of freedom and responsibility argue that a “true” freedom of the will – one which allows an agent to take advantage of genuine alternative possibilities and yet retain ultimate control over her decisions and actions – *does exist* but is incompatible with the deterministic picture that the future is causally necessitated by the past. Despite having an arguably intuitive appeal (and perhaps even exemplifying what many would instinctively take to be needed for any kind of *real freedom*), the libertarian position has received a good deal of criticism throughout the long history of the debate with many theorists believing the entire concept of an indeterminist freedom to be simply incoherent. Famously called “obscure and panicky metaphysics” by P.F. Strawson, “confused and empty words” by Thomas Hobbes, “the best self-contradiction that has been conceived so far” by Friedrich Nietzsche, and “moral levitation” by Daniel Dennett;[1] libertarian theories are often considered to be mysterious, counterintuitive, and even entirely incomprehensible and thus – critics believe – should simply be discarded. It remains a prominent and attractive position, however, and a good number of philosophers are committed to discussing and developing libertarian models. In addition, researchers working in other academic disciplines (such as neuroscience, physics, and mathematics) have also joined the debate on the side of libertarianism, apparently spurred on by seemingly indeterministic developments in fundamental physics as well as our growing understanding of the structure and function of the brain.

The question is whether the contributions of these scientists and philosophers supportive of the libertarian approach have moved us any further towards the development of an intelligible theory or whether the oft-cited problems – such as incoherence, inescapable arbitrariness, and plain mysteriousness – will prove to be fundamental to any indeterminist model. In what follows I shall attempt to answer this question and will ultimately argue that libertarian theories (or at least those reviewed in this thesis) *cannot* be considered fully coherent, as judged by the standards I set out below. To begin with, though, I will briefly outline the main challenges and key objections that libertarian theories face and in doing so clarify the reasons why they are often considered to be incoherent in the first place.

[1] Strawson, PF. 1962, p25. Hobbes – as quoted in Chappell (Ed.) 1999, p.73. Nietzsche 1889 (as quoted in Strawson, G. 1994, p.15). Dennett 2003, p.101.

1.1 The Libertarian Challenge

It should be noted from the outset that the terms “determinism” and “indeterminism” are not as straightforward as they are often taken to be and the meanings and implications of both concepts – and how they are employed by the researchers whose work I include in this thesis – will be considered more fully in the final sections, but broad definitions can be given as follows. Determinism is the thesis that all events have sufficient antecedent causes; with the inference being that, given a particular state of the universe at a particular point in time, there can only ever be *one* possible future from that point forward which is dictated by the conjunction of this past physical state and the relevant natural laws. Indeterminism, by contrast, is the thesis that not *all* events have such sufficient antecedent causes and, as such, there is genuine openness.

Libertarian theories fall on the indeterminism side of the argument and are taken to be those based on the following premises:

1) The only acceptable form of freedom is a specific type of free will that is incompatible with determinism.
2) This type of free will exists.

The goal for the libertarian theorist is thus to reconcile their understanding of free will with indeterminism: meaning they need to provide a positive account for how an agent can act freely and with sufficient control (both of which are key in order to ascribe any form of responsibility) in an indeterministic universe. There are of course numerous compatibilist arguments which attempt to show that freedom can be reconciled with *determinism* (and these will be discussed further in the final sections) but, for libertarians, the version of freedom which the compatibilists are arguing for is normally not considered to be the “Gold Standard” they themselves are seeking and believe exists. The perceived lack in a deterministic universe of what has been called “deep openness” by Alfred Mele (Mele 2014, p.2) – the ability to act differently given the exact same past conditions – leads to accusations that compatibilist agents are limited only to exercising “freedom of action” and thus do not possess a true “freedom of the *will*.” The former (somewhat broadly defined) I take to be the ability of the agent to act freely (in the manner of their choosing) without suffering any proximal impediments or constraints, such as being physically coerced, under threat of violence or under the influence of drugs and so forth. The latter (again, broadly considered) I take to add the further condition that the agent not be (wholly) subject to any form of necessitating causes over which they can have no control, whether that be the motions of atoms or the mechanical output of their own cognitive processing. Thus, to attain the Gold Standard freedom of the will, an agent requires the ability to carry out (at least some) actions or make (at least some) decisions that are not wholly determined by prior states of the universe, but which are nonetheless sufficiently within their control.

Many compatibilists (as well as others such as hard determinists or hard incompatibilists/impossibilists)[2] retort that this libertarian notion of "free will" is simply nonsensical and argue that any attempt to formulate a theory based on such principles can only end in failure for reasons which generally include the following:

1. Decisions or actions which are *not* wholly determined by past conditions in conjunction with natural laws can inevitably only be the result of *random* processes and thus can never be truly under the control of the agent.
2. Any attempts by libertarians to ascribe a sufficient form of control to the agent by firmly rooting their decisions and actions within a prevailing character (or *will*) presumably must involve some level of determination by this character, but this then raises questions over the formation of that character and whether the agent themselves can actually be responsible for forming it.
3. Attempts to address points 1 and 2 (as well as counter other criticisms and objections) usually leave libertarians trying to account for free will by invoking obscure or mysterious forms of agency and/or causation which, as well as often being difficult to comprehend, can bring with them ontological commitments that are difficult to accept.

Point 1 is a particularly common attack levelled at libertarian theories and leads to the argument that the required indeterminism – far from providing the agent with a necessary, unrestricted universal framework within which she can act freely, responsibly and with control – is actually destructive and serves only to *diminish* her control because choices are ultimately either determined or they are not, and undetermined choices, by definition, cannot truly be under the control of anyone or anything.

[2] Hard determinists are those who agree with libertarians that free will is incompatible with determinism but argue that determinism is true and thus free will does not exist. Hard incompatibilists (or impossibilists) are those who argue that free will cannot be reconciled with either determinism or indeterminism and thus cannot exist at all.

In regards to point 2, it would indeed seem that if we are to make any sense of the notion of a decision or action being fully within an agent's control – of a choice being ascribed to them in a manner in which we would consider them to be responsible for it – such decisions must somehow be firmly rooted within that agent's character or, to be more specific, must be the product of reasoning processes derived from their accumulated mental content (the sum of their nature, experiences, learning, desires, beliefs, proclivities and so forth), which I shall term their *Current Established Psychology* (CEP). However, introducing an essentially determining psychology (even if considered as only a "local" incidence of determination which is in no way indicative of any wider deterministic framework) in order to provide a sufficient form of agential control inevitably raises questions concerning where the responsibility lies for the formation of such psychology; setting off a seemingly infinite regress of a CEP formed by decisions made by a previous version, which was in turn formed by decisions made by a previous version, and so on back to a point before any rational decisions could possibly be made.

So, (according to the critics of the libertarian position) the positing of decisions and actions undetermined by past conditions leaves such decisions and actions overly subject to random factors which cannot be under the control of the agent, thus negating genuine agential responsibility. Any additional positing of some form of determining CEP in order to stave off any accusations of randomness likely leads to a problematic regress concerning the formation of the character from which an agent's decisions and actions would flow, which *also* seems to negate any true agential responsibility. Any further attempts to halt such a regress by specifically utilising the indeterministic foundations to posit causal breaks at specific points in the various chains of reasonings, decisions and actions just reintroduces the random factors, *once again* negating any true agential responsibility.

This dilemma speaks to the heart of the coherence question and is what I shall refer to as the ***Catch-22 of Libertarianism***. Its intractability is often what has caused libertarian theorists to fall foul of point 3.

1.2 Two Key Objections

Two specific objections have been formulated out of the more general criticisms above, which are intended to decisively refute all libertarian theories. They have been presented in a number of different guises from various theorists but have come to be known predominantly in the literature as the Luck

Objection and the Basic Argument; with the former focussing on a perceived inherent randomness (or chanciness) in indeterminist theories and the latter highlighting apparently inescapable contradictions that arise when an agent, in order to be free, is required to be (at least in some sense) *causa sui:* the cause of herself. Both objections are key to the question of coherence, and both will therefore need to be countered in order for a libertarian theory to have a chance of being considered successful. As such, both will feature heavily in the forthcoming discussion.

1.2.1 The Luck Objection

The Luck Objection is born out of one of the key tenets of libertarian theories of free will (alluded to in point 1 above), which is often labelled the “Alternative Possibilities” (AP) condition (it is sometimes also referred to as the “Principle of Alternative Possibilities” or the “Could Have Done Otherwise” condition). This condition essentially refers to the requirement that, at any point in time, an agent must be free to *act* or *act otherwise* whatever the past conditions and natural laws. Or, to put it another way, an agent must have been able to have acted otherwise at a particular point in time, *all past circumstances being exactly the same*. This principle is seen by libertarians as essential to prevent an agent’s decisions and actions from being wholly determined by antecedent events; such determination, they believe, being incompatible with their Gold Standard freedom of the will (and this being their predominant reason for rejecting compatibilism). The problems arise when we consider the potential implications of this AP condition: that if indeterminism means agents can genuinely have different alternatives open to them that all still remain consistent with their past and the laws of nature, it could be argued that we cannot conclusively account for why one alternative is chosen over another. This leads some to the conclusion that the final choice can therefore only be a matter of luck.

Alfred Mele, who has put forward one much discussed recent formulation of the objection, puts the problem as follows:

> If there are causally undetermined or indeterminate aspects of a process that terminates in, for example, a choice and these aspects are present at the very time the choice is made and directly relevant to the process's outcome, then, to the extent that the agent is not in control of these aspects, luck enters the picture in a way that threatens both moral responsibility and a desirable freedom. (Mele 1998, p.582)

Consider the following example. Jessica is a professional athlete on her way to a qualification race meet where, if she wins, she will qualify to compete in the next Olympic games, likely changing her life greatly for the better. She is running late and knows that if she misses the race her dream will be over. She comes to a crossroads where she intends to turn left and stops at a red light. Whilst waiting for the light to change, she notices what appears to be an elderly man walking away from her some way ahead up the road. Just at the moment the light turns green, the man seems to collapse to the ground. Jessica freezes and looks around desperately, but it is early in the morning and she can see no one else in the street. The man remains lying on the ground, with the reason for his fall or his current physical condition impossible to tell from her position. Naturally, Jessica enters into a state of inner conflict which reflects her competing desires and which draws on her existing nature, current reasons and beliefs, past experiences, and the other mental content and characteristics that comprise her CEP, which will ultimately decide her action. She may well feel instinctively that she should do something to help the man (indeed that she has a moral obligation to do so) but she may also acknowledge that in doing so she would likely miss out on her lifelong dream and all the potential glory that might follow. Jessica's hand moves to the door handle, poised, ready to open it and get out of the car, but then she returns it to the steering wheel, turns left and drives on to her race.

Consider now that we were able to wind the universe back to the precise moment when Jessica has her hand on the door handle, poised for action but not committed to it. Libertarian theories, based as they are on indeterminism and subscribed as they are to the AP condition, say that at the point just prior to committing herself to one course of action Jessica must be free to act otherwise, given the exact same past circumstances. Thus, the exact same past experiences, recent deliberations and CEP that resulted in Jessica driving on, could still have resulted in her getting out to assist, which is now what she does. Wind back the world once more and, perhaps, she drives on – and so on and so forth.[3]

The Luck Objection, then, proceeds as follows: if it is equally possible for the agent to act or act

[3] This example is a version of van Inwagen's "Roll-back" argument (see van Inwagen in Kane (Ed.) 2002, p.171). I in no way claim it to be particularly original – and many similar ones have gone before – but the protagonist is deliberately portrayed in such a way which gives her strong competing desires and reasons to act both ways. Mele's own example to illustrate the Luck Objection uses the idea of an agent in two possible worlds, which are the same up until final moment of decision/action, and where one agent does one thing and the other something different, but the core notion is the same (see Mele 1998, p.582).

otherwise given the exact same past circumstances, then there seems to be nothing within the agent (within Jessica herself) that is the deciding factor between her making the choice to stop and help the man and making the choice to drive away. Indeed, there is no deciding antecedent factor *at all* and so the actual outcome – whether she stops to help or not – is just a matter of luck. The Jessica in the first history (luckily or unluckily depending on your point of view) failed to overcome the temptation to put herself first and did not help the man. The Jessica in the second history succeeded in quelling her selfish urges and helped. Our instinct may well be to see the first Jessica as behaving immorally and the second as behaving morally, but if both Jessicas' backgrounds, states of mind, characters, motivations, deliberations and so forth (their entire CEP) are all *exactly the same* right up to the point where she either helps or does not help, it is argued that the first Jessica was just *unlucky* not to have stopped to help. It is therefore difficult to see how she could have had ultimate control over the decision in such a way as would be required if we are to consider her responsible for it.

This objection has been very difficult to dispel and, despite a number of libertarian theorists attempting to show how it can be countered (as we shall see in the forthcoming sections), it remains a prominent (negative) force in discussions of indeterminist theories.

1.2.2 The Basic Argument

The second key objection is born out of one of the other main tenets of libertarian theories of free will (as alluded to in point 2 above) and is one featured particularly in the work of the philosopher Robert Kane whose model is discussed in detail below (section 2.3) – the requirement for "Ultimate Responsibility" (UR). In short, the UR condition states that if an agent is to be ultimately responsible for any act, she must also be responsible for any preceding actions, decisions or events which can be said to be causally responsible for that act. Or, to put it another way: if an agent is to be considered fully responsible for acts dictated by her *will* (by her character or CEP), she must be fully responsible for the *formation* of the will from which such acts flow. The problems which arise in consideration of this condition have resulted in many believing free will (in the libertarian sense) to be simply impossible *whatever the truth of the fundamental nature of the universe, be it deterministic or indeterministic*. Such a view is encapsulated in the Basic Argument, an old argument recently formulated in the most developed way by Galen Strawson,[4] which focusses on the (proposedly incoherent) notion that a human being could be *causa sui*: the cause of oneself. The overriding claim of this argument is that we cannot ultimately choose what we do because we cannot ultimately choose *who we are*. We may act freely – be able to carry out "free actions" – in the sense (as discussed above) that we are not constrained by any obvious external forces, such as mental health conditions, addictions, coercive captors and so forth, but if we are not ultimately responsible for the formation of our CEP from which our actions arise then we still would not have the Gold Standard freedom of the *will* that libertarians are after.

Strawson presents various different versions of the argument in a number of different publications, but the following I take to be one of the clearest presentations of the argument's strengths:

> (1) You do what you do because of the way you are.
>
> So
>
> (2) To be truly morally responsible for what you do you must be truly responsible for the way you are - at least in certain crucial mental respects.
>
> But
>
> (3) You cannot be truly responsible for the way you are, so you cannot be truly responsible for what you do.
>
> Why can't you be truly responsible for the way you are? Because
>
> (4) To be truly responsible for the way you are, you must have intentionally brought it about that you are the way you are, and this is impossible.
>
> Why is it impossible? Well, suppose it is not. Suppose that
>
> (5) You have somehow intentionally brought it about that you are the way you now are, and that you have brought this about in such a way that you can now be said to be truly responsible for being the way you are now.
>
> For this to be true
>
> (6) You must already have had a certain nature N in the light of which you intentionally brought it about that you are as you now are.
>
> But then
>
> (7) For it to be true you and you alone are truly responsible for how you now are, you must be truly responsible for having had the nature N in the light of which you intentionally brought it about that you are the way you now are.

[4] See Strawson, G. 1986, 1994 and 2002.

> So
>
> > (8) You must have intentionally brought it about that you had that nature N, in which case you must have existed already with a prior nature in the light of which you intentionally brought it about that you had the nature N in the light of which you intentionally brought it about that you are the way you now are ...
>
> Here one is setting off on the regress. (Strawson 1994, p.13)

Utilising my own terminology, then. We act as we do, in the present, due to our current established psychology (CEP) – the way we are now, mentally speaking – as developed by our heredity, choices and experiences. If we are to be genuinely responsible for how we act in the present, we must be ultimately responsible for the *formation* of our CEP. But, in order to be ultimately responsible for the formation of our CEP, we would have to have carried out some actions in the past for which we can be held responsible which contribute to the formation of our CEP. But, in order to be responsible for *those* actions we must have been responsible for the formation of our established psychology (EP) that was in operation at that time in the past. But, in order to be responsible for the formation of our EP that was in operation at that time in the past, we would need to have been acting responsibly at an earlier time still, which would contribute to the formation of this past EP. And thus begins a regress. So free will is impossible because, ultimately, we cannot be the cause of ourselves.

Returning to the example of the athlete Jessica and her Olympic ambitions, we can once again picture her with her hand on the door handle, poised to act, and then imagine she returns it to the steering wheel and drives away leaving the elderly man to his fate. We could add that the man, elderly and frail as he is, actually fell because he had just suffered a cardiac arrest and, sadly, he died right there in the street. Jessica – a lifelong participant of athletics clubs – has perhaps been regularly first aid trained and therefore knows how to perform CPR. Many would instinctively say that, under such circumstances, she should have stopped to help and has therefore acted immorally by driving away without intervening. Though clearly not directly responsible for the man's death, it could be argued that by not stopping to check on him (and then likely using her skill to try and save his life) she shares some form of culpability. But, in light of the foregoing discussion, we must now consider on what basis can such conclusions be drawn.

Following the Basic Argument, in order for Jessica to be responsible for actions which were ultimately decided upon due to her CEP, she would have to be responsible for the formation of that CEP. But her CEP has been formed (at least in part) by her past actions and decisions, for which she would also need to be responsible. Let us say that her drive to succeed in her sport on the largest stage, which has ultimately contributed to her decision to abandon the elderly man and drive on to her qualification race, has its roots partly in her decision to join the athletics club three years earlier whilst at university. In order for her to be responsible for *that* decision (the decision to join the club) she would have had to be responsible for even earlier actions that contributed to the formation of her EP in operation at the time the decision to join the athletics club was made. Let us say that her interest in athletics, which would later result in her joining the athletics club at university, began when her father (himself a keen runner) enrolled her in a junior athletics programme when she was just five years old and, from that day forward, passionately encouraged her to compete. Would anyone wish to lay the responsibility for the decision to join the junior programme on the shoulders of a five-year-old Jessica? Perhaps not many would and, even if she were held responsible, we could just continue back even further.

The strength of the Basic Argument, then, relies on the seemingly intuitive picture of the development of the human psyche as a continual chain of actions and experiences from some early point in an agent's development (when her actions would be directed solely by conditions over which she can have no control, such as genetics and very early experiences) right up to the moment in which her adult CEP determines her current decision. It could be argued that all we need do to defeat it, therefore, is break the chain by claiming that our actions are not *always* determined by our CEP, or that its formation is not *wholly* reliant on our past decisions and experiences – perhaps by introducing indeterministic factors. As I have just shown in my discussion of the Luck Objection, however, if we do follow such paths we risk being led straight back to the problem of luck, with our decisions and actions potentially not being caused by *us* and thus not under our control. If a particular course of action is not wholly decided upon by something within us, such as our CEP, or it is but *we* are not wholly responsible for the formation of said CEP, then – it could be questioned – can we really say it is *our* action at all?

There is clearly a great deal more to say on such issues (I shall be discussing them at length throughout this thesis, and in the concluding sections in particular) and many philosophers and scientists do offer a defence of their position against both the Basic Argument and the Luck Objection. I shall cover the potential rebuttals in greater detail as the discussion progresses.

1.3 The *Criteria for Coherence*

The above discussion highlights the arguably intractable problems for libertarian theories of freedom and responsibility. On the one hand, the Luck Objection appears to show that any posited breaks in the causal chains to prevent our decisions and actions being sufficiently caused by antecedent conditions leads to there being an unacceptable amount of arbitrariness present in the process. On the other hand, the Basic Argument claims to show that the positing of any kind of determining control would seemingly entail a determining psychology (CEP) that we could not possibly be responsible for creating, resulting in us not being ultimately responsible for those present decisions. Hence the *Catch-22 of Libertarianism*.

So difficult has it proven to satisfactorily overcome these objections, the history of libertarian theories is (as per point 3 above) littered with invocations of obscure, mysterious, and even mystical forms of agency or causation. As Robert Kane puts it:

> Libertarians have invoked transempirical power centres, nonmaterial egos, noumenal selves, nonoccurrent causes, and a litany of other special agencies whose operation were not clearly explained. (Kane 1996, p.11)

And this is precisely why antagonists of libertarian theories claim that indeterminist models can *never* be made intelligible, and thus the line of reasoning should simply be dispensed with entirely. As one prominent compatibilist (and libertarian critic), Daniel Dennett, puts it:

> Libertarians seem to think that you can have free will only if you can engage in what we might call moral levitation. Wouldn't it be wonderful to be able to levitate — and then to dash off in any direction with the merest flick of a whim? I'd love to be able to do that, but I can't. It's impossible. There are no such miraculous things as levitators, but there are some pretty good near-levitators: Hummingbirds, helicopters, blimps, and hang gliders come to mind. Near-levitation isn't good enough, though, for libertarians… (Dennett 2003, p.101)

The key question is, then: exactly what *would* it take for a libertarian theory to be considered fully coherent? By which I mean, how could a theory successfully meet *all* the stated goals of the libertarian project in a consistent and non-contradictory fashion. Or, to put it another way: is there any way for a libertarian theory to escape from the ever-problematic *Catch-22 of Libertarianism*? In order to try and answer this question, I have developed a *Criteria for Coherence*[5] which includes six principles that identify the goals libertarians wish to achieve and which helps to crystallise the issues that seem to be at the root of any apparent incoherence.[6] It should be emphasised that it is my contention (support for which will be demonstrated throughout the discussion) that all six principles are explicitly endorsed by the stated aims and objectives of the philosophers and scientists whose work I shall analyse in the forthcoming sections. Thus, the standards against which I am judging their theories – and from which the *Criteria* has been derived – are the standards they have set for themselves and, as we shall see, the incoherence in their models arises as a direct result of the strategies they employ to try and meet them.

In the first instance, any purportedly libertarian theory will need to abide fully by the main tenets of the approach, such as denying the compatibility of the sought-after Gold Standard freedom of the will with determinism yet affirming its existence and therefore its compatibility with indeterminism. Thus:

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.
P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.

Next, the most common criticism that libertarians level at compatibilist theories is that determinism ultimately only allows for a particular future necessitated by past physical conditions. An agent in any given circumstance, they argue, would therefore only ever be able to follow *one* course of action (the course of action she ultimately does follow), which is prescribed by the past state of the universe in conjunction with the relevant laws of nature, and all other options which appear open to her are, in actuality, only illusory. If that is their charge against compatibilism, then a libertarian theory must (by contrast) fully meet the AP condition and allow for the existence of what they would consider "genuine alternatives." Thus:

[5] In the interests of clarity, the full appellation *Criteria for Coherence* is intended to be read in the singular, denoting the full set of principles as a whole, as is the single word *Criteria* when used as shorthand for the full appellation. At such times, italics will be used to avoid any confusion.

[6] Once again, I make no claim to huge originality with the individual principles proposed here, only that their drafting together to form a list of conditions against which libertarian theories can be formally assessed is a useful step.

P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.

The following two principles address key concerns of those critical of the libertarian position and are integral to the question of whether libertarian theories can ever be considered fully coherent. Firstly, indeterminist theories are seemingly open to accusations (as raised most directly by the Luck Objection) that they cannot adequately account for a sufficient agential control (and thus a justified ascription of responsibility), as undetermined acts cannot – by definition – be controlled by anything at all. A libertarian theory therefore needs to demonstrate a form of undetermined (*non*-determining) control which could also not be said to be solely down to luck. Thus:

P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.

Secondly, (as put forward in the Basic Argument) any attempt to avoid complete arbitrariness by positioning agential responsibility fully within an agent's CEP (or character, or *will*), which has been formed over the course of a life, can seemingly only lead to an infinite regress. An agent in an indeterministic universe may well be free to do and do otherwise but, in order for her to have control over what she ultimately does (freely) choose, the decision still needs to come directly from within her character. But if that is the case, who is responsible for the formation of that character? A theory therefore needs to demonstrate how (in at least some sense) an agent can be the cause of herself. Thus:

P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.

Finally, as discussed above, it is a common criticism of libertarian theories that they are overly obscure or mysterious. Thus:

P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

Given that these are the standards libertarian theorists have specifically set for themselves, it is my contention that only those theories that can fully meet these principles are ones which their authors can claim to be entirely coherent.

1.4 The Route to Progress

The format for this thesis will be as follows. In Part One I shall review the three main libertarian approaches, as differentiated in the philosophical literature: event-causal approaches (sections 2.2 and 2.3), non-causal approaches (sections 2.4 and 2.5) and agent-causal approaches (sections 2.6 and 2.7), as presented in the work of philosophers Robert Kane, Carl Ginet and Timothy O'Connor, respectively. The main theories of each will be presented and then assessed against my *Criteria for Coherence*. Each will be found not to have adequately met all my criteria meaning they cannot, by this standard, be considered coherent.

I shall turn in Part Two to the work of researchers from other academic fields who have specifically turned their attention to the philosophical problems of free will in their research to assess whether they can claim any more success in developing the kinds of libertarian theories which could meet the *Criteria*. The three theories I shall review are: Criterial Causation from Neuroscientist Peter Ulric Tse (section 3.2), Orchestrated Objective Reduction from mathematician Roger Penrose and biologist Stuart Hameroff (section 3.3) and Quantum Interactive Dualism from physicist Henry Stapp (section 3.4). Once again, the theories will be assessed against the *Criteria* and I will show that each of them also fails to adequately adhere to its principles.

In the Conclusion I will briefly summarise the ways in which all the approaches considered have failed to meet the *Criteria* and discuss how analysing the theories against these six principles illustrates, first and foremost, the scale of the task libertarian theorists have set themselves. Furthermore, I will show that the root cause of the incoherence is the very feature that makes libertarian theories *libertarian*: namely, the obligatory attempt to explicitly incorporate indeterminism in their models for free and responsible actions. I will go on to consider whether indeterminism really is required in order to meet both the AP and UR conditions and ultimately argue that, at least for all the approaches considered, the indeterminism which is portrayed as essential to the theory is, in fact, broadly irrelevant to the model of agency being described. Finally, I will question the usefulness of the current framing of the debate as a “contest of compatibility” between free will and determinism, on the one hand, and free will and indeterminism, on the other. I will argue that this long-standing approach is mistaken and suggest a different path that could perhaps now be taken to move the debate forward.

Part One

The Philosophers

2.1 Introduction

Libertarian theories of free will fall mostly into three different camps: event-causal, non-causal and agent-causal approaches.[7]

Event-causal theories (as the name suggests) focus on indeterministic causation by events internal to the agent. Such theories often appear the more intuitively appealing of the libertarian approaches, due their straightforwardly materialist view of human agents and an arguably *prima facie* instinctive attractiveness. The major obstacle for such approaches, however, is the core concept of indeterministic (undetermined or non-determining) causation, as many question the very coherence of this entire notion. Employing it, as we shall see, leaves the theories vulnerable to attack from both the Luck Objection and the Basic Argument, as well as a target for accusations they can only offer diminished agential control.

Non-causal theories (again as the name suggests) try to remove problems rooted in determining, antecedent causation by denying the very fact that the agent need *cause* anything at all. It is believed by theorists developing such an approach that, by removing the requirement that an initial (and initiating) mental action or volition be caused in any way by the agent, we can remove any problematic regress yet retain a required level of agential control. Breaking the causal link in such a comprehensive way, however, leaves such theories once again susceptible to both the Luck Objection and the Basic Argument, with (arguably) the additional detraction of making a particularly unconvincing case for a sufficient form of control. Furthermore, it is argued that the theories involve what many view as an obscure approach to both the motivations for, and the mechanism of, human action, as well as a mysterious approach to causal notions in general.

Agent-causal theories present the agent as an enduring *substance*, which directly causes its actions (or the formation of intentions to act) in a manner not reducible to events of any kind. Though enjoying somewhat of a recent resurgence in the work of philosophers such as Timothy O'Connor and Randolph Clarke, such theories have been largely out of favour in modern times. This is due mostly to the

[7] The first two types of libertarian theory are sometimes referred to as Causal-indeterminist theories and Simple-indeterminist theories, respectively.

necessary positing of an enduring agent-as-substance, which some have argued has mysterious origins and constitution; one which appears in some sense outside of the causal order. Though, *prima facie,* seeming to offer *some* defence against the Luck Objection, it suffers much the same problems in the face of the Basic Argument as the other libertarian approaches and suffers the added charge of invoking an obscure entity that is the enduring agent.

I shall now review each type in turn and consider them against my *Criteria for Coherence*.

2.2 Event-Causal Theories

Event-causal (or causal-indeterminist) approaches focus on establishing a framework for indeterministic control by postulating agent-involving events which cause actions non-deterministically. Agents *cause* their actions in a seemingly understandable and linear way by having mental events (thoughts, feelings, memories, beliefs, desires and so forth) which cause other mental events (say, in a process of internal, reasoned deliberation) and ultimately cause physical events.[8] The agent can thus be said to be in *control* of their actions and decisions as these are caused, in an appropriate way, by their being in certain mental states. However, because they cause them non-deterministically there will always remain the possibility that a different decision could have been made and a different course of action followed. As discussed above in Section 1, this is essential for libertarians as they believe the presence of genuine alternative possibilities (thus meeting the AP condition) avoids our decisions and actions being wholly determined by prior conditions, which they consider would preclude an agent from achieving the Gold Standard freedom of the will they are striving for.

The event-causal account is therefore broadly similar to the compatibilist account, in that events within the agent (her reasons and so forth) *cause* subsequent behaviours, only with the one fundamental difference that they do so (at least on some occasions) non-deterministically instead of deterministically, and thus a different outcome is a possibility whatever the state of the universe at that time. Libertarians argue that only this indeterministic variant of mental processes would allow – *contra* compatibilism – for *genuine* consideration of all physically possible eventualities (and thus ultimate responsibility, UR) because the eventual outcome is not determined by the prevailing physical conditions. The indeterminism is what allows for free *will*, rather than just an agent's ability to carry out free actions. A key question is whether we can really make sense of such a concept as indeterministic causation and whether, by using it as a platform for autonomous action, we can ever attribute sufficient agential control, avoid falling foul of the Luck Objection and/or the Basic Argument and escape from the *Catch-22 of Libertarianism*.

[8] This distinction between mental and physical events is in no way meant to imply that mental events are not physical events, it is simply phrased in such a way as to delineate mental events, such as our thoughts and feelings, from other types of events.

A number of different variations of the event-causal approach have been proposed, with disagreements focused largely on exactly *when* in the extended causal process (which potentially spans the entire life-history of the agent) the required indeterminism should feature. One of the most developed and widely discussed versions – and one which directly confronts both the Luck Objection and Basic Argument – is the account advanced by Robert Kane which requires that (at least some) free actions are *themselves* indeterministically caused. I will now consider Kane's account in depth, on the understanding that the analysis of his model and its assessment against my *Criteria for Coherence* likely highlights what conditions *all* such views would need to meet to be considered coherent.

2.3 Robert Kane's Self-Forming Actions

Robert Kane's work in this area is considered by many to be the best attempt to present a coherent libertarian theory of free will. He seeks first and foremost to challenge the common assumption that the mere presence of indeterminism means things can *only* happen by chance or luck and therefore, when applied to the acts of an autonomous agent, would likely preclude any form of responsibility. Indeterminism, Kane argues, just means that outcomes are not necessarily determined by antecedent causes and nothing further should be automatically inferred. Secondly, he wishes to avoid any contentious ontological additions, unnecessary obscurity or extraneous mysteries and seeks in much of his work to underpin his model with solid physical and neurological foundations in order to demonstrate that libertarian theories can be examples of rigorous, scientifically informed philosophy. As such, he attempts to distance himself from other libertarian theorists who have chosen to pursue what he considers to be ontologically extravagant "extra-factor" approaches that usually require the introduction of obscure forms of agency or causation; introducing what he terms "The Free Agency Principle":

> In the attempt to formulate an incompatibilist or libertarian account of free agency that will satisfy the plurality conditions and UR, we shall not appeal to categories or kinds of entities (substances, properties, relations, events, states, etc.) that are *not also needed by nonlibertarian* (compatibilist or determinist) *accounts of free agency satisfying the plurality conditions.* The only difference allowed between libertarian and nonlibertarian accounts is the difference one might expect—that some of the events or processes involved in libertarian free agency will be indeterminate or undetermined events or processes. But these undetermined events or processes will not otherwise be of categories or ontological kinds that do not also play roles in nonlibertarian accounts of free agency (such as choices, decisions, efforts, practical judgments, and the like)—the difference being that in nonlibertarian theories, these events or processes need not be undetermined. Such differences as there are between libertarian and nonlibertarian theories should flow from this difference alone, and the task will be to make sense of a libertarian freedom satisfying the plurality conditions, given this difference. (Kane 1996, p.116)

I will return to what he means by the plurality conditions and UR shortly in this section, but it is clear that Kane – whilst pitching his tent firmly in the libertarian camp – is not looking to hide from the evident problems with the indeterminist approach and is also not intending to seek the kinds of support which would most likely be unavailable to his opponents.

2.3.1 AP vs UR

Prior to Kane's work incompatibilist intuitions were driven largely by the focus on the availability of genuine alternative possibilities (on how to meet the AP condition) – the requirement that the agent "could have done otherwise" – with the main emphasis placed on demonstrating that this condition is incompatible with determinism. By combining AP with indeterminism, Kane formulates what he terms the "Indeterminist Condition," which coins a principle generally viewed by libertarians as integral to achieving the Gold Standard free will:

> *The Indeterminist Condition*: the agent should be able to act and act otherwise (choose different possible futures), *given the same past circumstances and laws of nature.* (Kane 2005, p.38)

The truth of this condition (a version of which constitutes principle three in the *Criteria*) is important for libertarians because it ensures that whatever choice an agent finally makes, or whatever action she ultimately performs, is not necessarily determined *at any stage* by prior states or events – affording the "deep openness" they hold dear. For Kane, however, the reason the debate stalled for incompatibilists was because AP was being considered in isolation and such a strategy could never resolve the question of whether free will is compatible with indeterminism because it leaves out an important (perhaps the most important) factor when considering the capabilities of autonomous agents. If we can be said to be *freely choosing* between available options, the sources or origins of these choices (and thus any subsequent actions) must be "in us" rather than in something else, such as the laws of nature working on antecedent conditions and so forth. For Kane, therefore, it is actually the less debated condition of Ultimate Responsibility (UR) that is the key to building the case for a libertarian position, and only by focussing on a combination of AP, indeterminism and UR can we hope to make progress towards a coherent theory.

Kane defines UR as follows:

> (UR) An agent is *ultimately* responsible for some (event or state) E's occurring only if (R) the agent is personally responsible for E's occurring in a sense which entails that something the agent voluntarily (or willingly) did or omitted, and for which the agent could have voluntarily done otherwise, either was, or causally contributed to, E's occurrence and made a difference to whether or not E occurred;

> and (U) for every X and Y (where X and Y represent occurrences of events and/or states) if the agent is personally responsible for X, and if Y is an arche (or sufficient ground or cause or explanation) for X, then the agent must also be personally responsible for Y. (Kane 1996, p.35)

Agents, therefore, can be ultimately responsible for an action only if: a) they performed the action; b) they could have done otherwise than what they did; and c) they can also be considered the *ultimate source* of the action. For c), they can be considered the ultimate source only if they are also responsible for any preceding actions/decisions which caused the action in question or for forming the motivations or intentions from which the action/decision results. In other words, the agent can be the ultimate source of their actions only if they are responsible for the formation of their current established psychology, CEP (or character, or *will*) from which their actions flow. Free will, Kane is arguing, is about more than just freedom of action – more than merely having the freedom to act and act otherwise – it is about the agent's role in the development of their character, in them becoming the person they are, which is ultimately the source of their actions whether they be praiseworthy or blameworthy. Thus, while indeterminism and AP are *necessary* for free will they are not *sufficient*.

One way Kane seeks to demonstrate this point is by showing how we can readily conceive of instances where a person could indeed "do otherwise" and their actions be undetermined, and yet apparently still lack free will. An example he gives to demonstrate this (which he bases on the work of J.L. Austin)[9] involves a rooftop assassin, armed with a rifle, who is trying to kill the Prime Minister. The assassin takes his shot but, due to a random nervous twitch in his arm at the crucial moment, he misses his intended target and kills the Prime Minister's aide instead.[10] The occurrence of the twitch and (more importantly) the potential for such an interfering, neurological symptom to occur completely at random makes the killing of the Prime Minister a genuinely undetermined event, satisfying the indeterministic requirement for libertarian free will. Also, as the outcome of the shot is genuinely undetermined, he might well have succeeded in killing his intended target. Indeed, as an expert assassin, he has made many similar shots before and so has the skill and the opportunity to succeed. Thus, the assassin could have done other than miss his target. We therefore seem to have an action which is both undetermined and such that the agent could have done otherwise but where we question the freedom of the protagonist as the

[9] See Austin 1956, in which he discusses a golfer missing a putt due to a twitch in his arm.

[10] Kane 2005, p.124, and similar in Kane 1996, p.110.

assassin did not miss his target *voluntarily* – missing was out of his control, despite the fact that missing was something he *did* (even if by accident).

Kane moves to strengthen this argument and its importance for the wider notion of free will by inviting us to further imagine that external influences are interfering with how our CEP is configured. Let us imagine that a nefarious neuroscientist has somehow implanted the relevant intentions, reasons, beliefs, desires and so forth in the mind of the assassin who concretely sets his will to accomplish his task.[11] Now, whether he makes the shot or not, his intent is set and has been determined by the neuroscientist. Kane argues that the assassin in such a scenario would completely lack free will, even though the act is undetermined and there are alternative possibilities available, because they can only do otherwise in the limited way discussed by Austin: they can only do otherwise accidentally or unwillingly. They may perform the action but are not responsible for the *will* that drives it, and thus could not be considered *ultimately responsible* for the action. Hence the importance of UR as well as AP and indeterminism.

2.3.2 Kane's Plurality Conditions and the "Problem of Plurality"

The above discussion draws out key features of the libertarian approach and highlights the issues that arise when we consider the importance of UR, AP, and indeterminism. Examples such as that involving the assassin seem to demonstrate that the featured indeterminism can only be shown to be compatible with the limited Austinian freedom that Kane labels *one-way rationality*[12] – where an agent's will is firmly set in a particular way and there is no further decision to be made and therefore he would "do otherwise" only inadvertently. If the assassin hits his mark, we are comfortable he is doing so by choice but, if he misses, he can do so only by accident. The assassin cannot both hit and miss voluntarily. Kane makes the point that were this concept of one-way rationality to be the benchmark for human free will, it would appear to many as a strange sort of freedom: one in which we can indeed act freely, rationally, and voluntarily but then only do otherwise *accidentally*. Furthermore, as the example of the nefarious neuroscientist showed, there is an additional problem with this kind of one-way rationality. As UR states, our actions (particularly any we would accept being truly responsible for) must come from *us*;

[11] Kane (2005, p.126) uses the example of the assassin's actions being predetermined by God but the outcome is the same.

[12] See Kane 1996, p.108 and Kane 2005, p.128.

from our CEP, character, or will. In the scenario we have painted above the assassin's will is set externally by a malign influence prior to the action and thus, if we are to assign (ultimate) responsibility, it would seem we must look back to the formation of such a will.

Kane argues, therefore, that what is required for genuine free will is not just *one-way* rationality but what he terms *plural* rationality: being able to both do or do otherwise voluntarily, intentionally, and rationally. Kane acknowledges, however, that by imposing a requirement for acts to meet such "plurality conditions" the case for free will being compatible with indeterminism becomes even more difficult to argue for as it leads to what he calls the *Problem of Plurality*, which breaks down as follows.[13] On a libertarian approach such as Kane's (in necessary compliance with the Indeterminist Condition), when presented with a choice to make, an agent needs to be able to choose any of the available options at any time – whatever the past circumstances – and do so rationally, without the choice being either determined or arbitrary (or indeed always accidental) *whichever option is chosen*. At no point in the deliberative process can we be determined by the past to make one choice over another at a given time, but the choice we do make must still be intentional and rational. The problem with this picture (and hence the problem of plurality) is that the notion of a choice which is completely under our voluntary control but entirely undetermined by past conditions and always subject to the possibility of sudden change – and yet not in any way random or arbitrary – seems to many to be at best confused and at worst entirely incoherent. The presence of this problem of plurality leads to the common charge that libertarian theories can only produce a counterintuitive kind of freedom based on luck and randomness.

Kane presents the following example to illustrate the problem. Jane is trying to decide whether to holiday in Hawaii or Colorado. There are lots of reasons to support each choice and, unlike our assassin, her will is set in neither way so a long and detailed deliberative process ensues leading her to favour Hawaii. However, if we accept the Indeterminist Condition, we have to accept that despite these deliberations (and given the exact same past circumstances and laws of nature) she could still have chosen Colorado, seemingly arbitrarily and out of the blue:

> This means that *exactly the same prior deliberation* up to the moment of choice, through which she came to believe that Hawaii was, all things considered, the best option, may have issued in the choice of Colorado – exactly the same prior thoughts and reasonings, the same imagined scenarios

[13] See Kane 1996, p.109.

> and considered consequences, the same prior beliefs, desires, and other motives that led to the choice of Hawaii – not a sliver of difference – would have issued in the choice of Colorado instead. This is strange, to say the least. (Kane 1996, p.107)

It seems difficult for Kane to account for this possibility without accepting that undetermined decisions under such circumstances can only be arbitrary. How can we rationally explain how an agent can be equally free to go either way at a decision point, despite all the past deliberations up to that point?[14]

All these considerations further reinforce the apparent *Catch-22 of Libertarianism*. On the one hand, one-way rationality (involving a pre-set will) can meet the AP and indeterminism requirements, yet not be sufficient for libertarian free will due to a lack of UR. On the other hand, plural rationality – which is required for UR – results in the problem of plurality; seemingly leaving our ultimate choices as solely and inescapably arbitrary. Kane's answer to this dilemma is to attempt to establish a model of libertarian free will that allows for ubiquitous one-way rational decisions, but which nonetheless can still be considered as truly *free* actions (in the Gold Standard libertarian sense) because they flow from a CEP that has been formed (at least in part) from key, defining – and probably far less frequent – plural rational decisions. This key addition is designed to bring UR back into the equation, as well as halt any potential regress. Precisely *how* Kane's model tries to achieve this is what I shall turn to next.

2.3.3 Self-Forming Actions

Emphasising the importance of UR allows Kane to make the important and (to many) appealing step to allowing that the AP condition need not necessarily be fulfilled for *every* act an agent does freely. We do not always need to be able to "do otherwise" at every relevant juncture in the plurally rational way in order to have free will, but we must have been able to do so for *certain* acts that we carried out in our lives that contributed toward the creation of our CEP/character/will from which the other acts (ones that do not meet the AP condition) result. Kane therefore allows that some acts are indeed sufficiently caused by prior events – by the agent's CEP – and, despite this apparent violation of the core

[14] It is important to note, of course, that were Jane to choose Hawaii at the end of her (indeterministic) deliberations that favoured Hawaii, this would not at all be surprising. It is the physical possibility of her ultimate choice being Colorado that causes the problem, which supports one of Kane's key points that indeterminism *per se* does not rule out any kind of free action for which an agent can be held responsible and he is explicit in arguing that we must, therefore, break the intuitive connection in our minds between "indeterminism's being involved in something" and "it's happening merely as a matter of chance or luck." (Kane 2011, p.391)

incompatibilist/libertarian principle, these acts can still be considered *free* (in the relevant sense) so long as the CEP from which they result has in its formative history some actions which were *not* determined by prior events and which conformed to the plurality conditions. Kane focusses on such instances where an agent does not necessarily have their will formed either way on a particular issue – where there is often competing and incommensurable courses of action open to them – and a free and undetermined action is required in order to settle the matter. He calls such actions "self-forming actions" (or sometimes "self-forming willings" to emphasise their "will-setting" nature):

> [SFAs] are the undetermined, regress-stopping voluntary actions (or refrainings) in the life histories of agents that are required if U is to be satisfied, and for which the agent is personally responsible in the sense of R. The agents must therefore be responsible for them directly and not by virtue of being responsible for other, earlier actions (as would be required if they were not regress-stopping). This means that, for SFAs, the "something the agents could have voluntarily done (or omitted) that would have made a difference in whether or not they occurred" is simply *doing otherwise,* rather than doing something *else* that would have causally contributed to their not occurring. (Kane 1996, p.75)

As the agent's will cannot be firmly set one way or the other on the issue for the decision to count as self-forming (and thus regress-stopping), she must be able to avail herself of the plurality conditions. She must, at this juncture, be able to act or act otherwise, and do both rationally, voluntarily and with control. Kane's SFAs, therefore, are different from the (potentially ubiquitous) actions and decisions which can flow determinedly from an agent's CEP and yet still be counted as free and ultimately undetermined acts (albeit only one-way rationally). The SFAs are the acts by which the agent over time freely *creates* her own will. If her will is *not* formed in this way but rather as just an accumulation of solely one-way rational acts, then no subsequent acts flowing determinedly from it could be considered free acts on Kane's account.

> The "self-forming actions" (or SFAs) required by UR must satisfy the plurality conditions. They must be more-than-one-way rational, voluntary, and voluntarily controlled. In the case of plural rationality or motivation, if this were not so, the wills of agents (their reasons, motives, intentions, etc.) would already be "set one way" when SFAs occur, and UR would require further backtracking to determine whether the agents were responsible for having the wills they do have. In other words, to play their buck-stopping roles, SFAs *must be "will-setting" and cannot presuppose a will already set.* Some free and responsible actions might be merely one-way rational, voluntary, and voluntarily controlled, but

> not all of them could be, if agents are to be ultimately responsible for forming their own wills. (Kane 1996, p.114)

The requirement for SFAs thus links together the three key concepts of UR, AP, and indeterminism. In order for an agent to achieve UR and thus be considered the creator of their own will in such a way as to permit moral responsibility, they must be able to avail themselves of these undetermined SFAs. As such, UR entails both AP and indeterminism. For the assassin, whose will is already set on murder, we can allow for one-way rational choices which can be reconciled with both AP and indeterminism, but without the addition of SFAs in his past no decisions would conform to the required plurality conditions and thus his murderous actions could not be considered free.

Jessica and the Fallen Man

Applying Kane's model to our conflicted athlete example allows us to further examine Kane's libertarian model. Jessica has two ways in which she can act freely.[15] As she sits, poised to act – perhaps with her hand on the car door handle – staring at the elderly man on the floor and worrying about missing the most important race of her life, her eventual decision could flow determinedly from her CEP and still be free (so long as this CEP has been formed partly by previous SFAs) or her decision could result from a new SFA. Given that the situation Jessica is facing can be considered the kind of typically "torn" dilemma which Kane believes most likely to involve an SFA, we shall assume Jessica's will is set neither way on the matter at the moment of decision.

Although the decision must remain undetermined right up to point when it is made, Kane would not wish this fact to imply that Jessica's CEP plays no role whatsoever in the production of the SFA. Presumably a multitude of considerations and calculations would rapidly filter through the relevant cognitive processing areas of her brain – whether consciously or unconsciously – as her deliberations progressed, but just not in such a way as to algorithmically churn out a decision. On the one hand she has a lifetime of hopes and dreams of athletic success (and perhaps also the accompanying financial incentives, which could well be desperately needed) and on the other the instinctive desire to assist a fellow human being who may be greatly in need. It is in fact such considerations in the face of two

[15] Kane's own example of a businesswoman on her way to an important business meeting who witnesses a robbery in an alleyway is similar (see Kane 1996, p.126).

incommensurable desires that, for Kane, generates the type of neurological activity that is required for the occurrence of an SFA, and it is the very conflicted nature of such neurological activity that generates the very indeterminism essential for libertarians.

> There is tension and uncertainty in our minds about what to do at such times, let us suppose, that is reflected in appropriate regions of our brains by movement away from thermodynamic equilibrium – in short, a kind of “stirring up of chaos” in the brain that makes it sensitive to micro-indeterminacies at the neuronal level. The uncertainty and inner tension we feel at such soul-searching moments of self-formation would thus be reflected in the indeterminacy of our neural processes themselves. What we experience internally as uncertainty about what to do on such occasions would correspond physically to the opening of a window of opportunity that temporarily screens off complete determination by influences of the past. (Kane 2005, p.135)

Kane calls such neurological activity “parallel processing” in the agent’s brain. It is a kind of mental partitioning, in which two different Jessica’s end up battling it out (metaphorically speaking). On this occasion, say, the Jessica who puts herself first wins the day and she hurriedly drives on to her race. Upon making this decision, a new SFA – one coloured by selfishness perhaps – is added to her CEP and her character could be altered forever (even if only marginally) by way of a selfish behaviour being reinforced or, conversely, altered in a different way by subsequent feelings of guilt and so forth.

An Initial Assessment

Whilst a detailed and useful model for undetermined decision making, it is not immediately clear how Kane’s SFAs provide any answers to the perennial problems of the libertarian approach, overcome his own “problem of plurality,” or tackle the *Catch-22 of Libertarianism* I have highlighted. There does not appear to be any definitive explanation for how these supposedly regress-stopping decisions – which require indeterminism, multiple options, plural rationality and must abide by the Indeterminist Condition – can be made without the result still being labelled as arbitrary. Jessica has two main competing options and has clear reasons for choosing either, so whichever choice she makes could indeed be said to be intentional, voluntary, and rational. However, given that these choices cannot be determined by her CEP, we must accept that the agent might choose either way *all past facts remaining the same*. Thus, that being the case, the question becomes: what explains the successful or failing effort to do the right thing? Or, in other words: what is it that (non-determinedly) tips the balance in favour of one option over another? The choice has to be influenced but not determined by Jessica's CEP or the libertarian case fails. But, if it is not determined by her character or motives, the possibility remains that – whatever the makeup of her CEP and whatever the resultant deliberations – at the end of the process she effectively has a clean causal slate to choose either option. It is strange to think that after weeks of deliberations strongly inclining Jane to favour Hawaii, she then suddenly books Colorado out of the blue with seemingly no reason for it. Without any prior state or event being allowed to tip the balance in any singular SFA, we seem to be slipping inexorably (and perhaps unacceptably) back towards luck and thus Kane needs to have an answer to the Luck Objection.

2.3.4 The Luck Objection

Given the foregoing discussion, Kane's theory would initially seem particularly vulnerable to the Luck Objection but he nevertheless attempts to tackle it head on in various works, often focussing on Alfred Mele's version of the objection,[16] which (adapted here to Jessica's dilemma) proceeds as follows. In the actual world Jessica chooses to ignore the plight of the elderly man, hurrying on to her race. If the Indeterminist Condition holds and she could have done otherwise given the exact same past circumstances, then her counterpart, Jessica*, in another possible world which is exactly the same as ours up to the point of decision, could have overcome her selfish urges and gone to help. Mele argues: "If there is nothing about the agents' powers, capacities, states of mind, moral character, and the like that explains this difference in outcome, then the difference is just a matter of luck" (Mele 1998, p.583).

Kane (1999, p.229/230) fully breaks down the strong version of Mele's argument as follows: [17]

a) In the actual world, person P does A at t, and does so voluntarily and intentionally.

On the assumption that the act is undetermined at t, we may imagine that:

b) In a nearby-possible world which is the same as the actual world up to t, P* (P's counterpart with the same past) does otherwise (does B) at t, and does so voluntarily and intentionally.

[16] See Mele 1998, p.582.

[17] The strong version excludes cases of simple one-way rationality and the possibility of doing otherwise accidentally, as in the case of Kane's assassin. I have taken the liberty of adapting Kane's "general form" of the argument on page 229 to include – in one place – his additions on page 230, which transform it into the strong version.

c) But then (since their pasts are the same), there is nothing about the agents' powers, capacities, states of mind, characters, dispositions, motives, and so on prior to t which explains the difference in choices in the two possible worlds.

d) It is therefore a matter of luck or chance that P does A and P* does B at t.

e) P is therefore not responsible (praiseworthy or blameworthy, as the case may be) for A at t (and presumably P* is also not responsible for B).

Kane believes this argument fails, but himself acknowledges it is not easy to show exactly *why* it fails. Recall that, unlike the assassin, Jessica's will is not already settled or formed on the issue that confronts her. Her motivations, desires, beliefs, memories, and all other relevant factors are feeding into her deliberations but, going into and during this process, her CEP is not configured such that either outcome is determined to occur, and neither will be until the moment a course of action is selected and embarked upon. However, despite this difference, it is indeed the assassin example (and other one-way voluntary situations) that Kane believes can show us a way around the Luck Objection.

As discussed briefly at the beginning of section 2.3, Kane argues that it should be clear to all that common usages of terms like "luck" and "chance" are not entailed by indeterminism alone and certainly do not apply to the assassin example in such a way as to reduce his culpability for his acts. The assassin could well miss his target, but would do so only by accident because his will is set to murder and no one would wish to say that him therefore succeeding in his murderous intent is only down to "luck" and he is thus not responsible. Kane argues that the same logic can be applied in the case of his SFAs. An agent in such a situation as Jessica's is in effect trying to follow *both* courses of action in the same way as the assassin is trying to follow *one* course of action so, for the same reasons as in that example, terms like "luck" and "chance" do not apply to her situation either and, as she is striving to achieve both outcomes, she can rightly be held responsible for which ever course of action is ultimately initiated.

Furthermore, for the assassin, the indeterminism represented by the potential arm twitch is essentially *external* to his CEP and can indeed be seen as being of the "hindrance" variety, which opponents of libertarian accounts often label indeterminism more generally. For the assassin, it is externally imposed and something to be *overcome* but during an SFA, to the contrary, indeterminism is in fact (according to Kane) *created by the neurological activity of the agent*. It is internal, not external – coming from *within* the will of the agent concerned.

> Imagine that in such conflicting circumstances, two competing (recurrent) neural networks are involved. (These are complex networks of interconnected neurons in the brain circulating impulses in feedback loops of a kind generally involved in high-level cognitive processing.) The input of one of these networks is coming from… [Kane's businesswoman's] desires and motives for stopping to help the victim. If the network reaches a certain activation threshold (the simultaneous firing of a complex set of "output" neurons), that would represent her choice to help. For the competing network, the inputs are her ambitious motives for going on to her meeting, and its reaching an activation threshold represents the choice to go on. Now imagine further that these two competing networks are connected so that the indeterministic noise that is an obstacle to her making one of the choices is coming from her desire to make the other. Thus, as suggested for SFAs generally, the indeterminism arises from a tension-creating conflict in the will. (Kane 2002, p.419)

Thus, Jessica's situation is a kind of "doubling" of the assassin example. The assassin's will is set to murder and he has to try and overcome the hindrance of some of the indeterministic factors involved (involuntary twitching and so forth) in the performing of the action in order to accomplish his task. He is trying to succeed and cannot both try to succeed and try to fail at the same time. If he succeeds, he is morally responsible for the murder regardless of the fact that he could have failed. If he fails, he is not responsible for the failure in the same way. Jessica, according to Kane, is effectively splitting her psyche into two competing parts (neural networks) as her will is not yet set. One part wants to stay and assist and the other wants to drive on to her race. Each part is then in the same one-way voluntary situation as the assassin and they vie with each other until one crosses the "activation threshold."

To analyse this further, let us call the two parts J1 and J2 respectively. J1 is striving to produce outcome A, J2 is striving to produce outcome B. As they are part of the same neurological self, the conflict stirs up indeterminism in the brain making the ultimate outcome indeterminate and uncertain. J1 is only striving for the moral outcome; J2 is only striving for the selfish outcome. Taken independently they would only be responsible for the outcome they were trying for, and failure to achieve the outcome could only be labelled as accidental, resulting in the rather limited form of freedom. However, taken as an integrated set (part of what Kane calls the "self-network"),[18] Kane believes we can ascribe responsibly to the unified Jessica *whichever way her decision falls*.

[18] See Kane 1996, p.137 for more on the self-network, which is similar in many respects to my CEP.

> The point is that in self-formation of these kinds (SFAs), failing is never *just* failing; it is always also a *succeeding* in doing something else we wanted and were trying to do. And we found that one can be responsible for succeeding in doing what one was trying to do, even in the presence of indeterminism. So… the [luck] argument [is] invalid for cases like the businesswoman's and other SFAs…" (Kane 1999, p.234)

Despite the eventual outcome of her decision being undetermined right up to the point the decision is made, Kane argues that terms such as "luck" or "chance" do not apply in Jessica's case and, moreover, they do not apply in the very same way in which they do not apply in the assassin case, because Jessica is *actively striving to achieve both outcomes*. She can therefore be held fully responsible for either choice and premise d) and e) of – Kane's version of – Mele's argument are shown not to follow. Furthermore, Kane adds the additional consideration that as Jessica has been actively striving for each outcome she would certainly consciously/psychologically *own* whichever choice she made as hers were she to be questioned about it afterwards, in a manner not consistent with the idea that the outcome was caused by some hindering neurological "epicurean swerve"[19] (to which undetermined decisions are often reduced by critics of the indeterminist position) or that her choice was simply down to luck. Appreciation of this phenomenal aspect of decision-making Kane believes to be essential to understanding his theory.

> If she [Kane's Businesswoman] succeeds in choosing to return to help the victim (or in choosing to go on to her meeting) (i) she will have "succeeded despite the probability or chance of failure"; (ii) she will have succeeded in doing what she was trying and wanting to do all along (she wanted both outcomes very much, but for different reasons, and was trying to make those reasons prevail in both cases); and (iii) when she succeeded (in choosing to return to help) her reaction was not "Oh dear, that was a mistake, an accident – something that happened to me, not something I did." Rather, she endorsed the outcome as something she was trying and wanting to do all along; she recognized it as her resolution of the conflict in her will. And if she had chosen to go on to her meeting she would have endorsed that outcome, recognizing it as her resolution of the conflict in her will. (Kane 1999, p.233)

[19] This notion of atoms swerving randomly was believed to be discussed by Epicurus in answer to the deterministic picture provided by Democritean atomism. The "swerve" was designed to introduce an uncaused cause of movement to halt the apparent vicious regress that would seemingly rule out any free will but was criticised for introducing only chance factors which could in no way enhance our freedom.

Kane thus appears to be proposing the following conditions, the meeting of which presumably results in a successful SFA and therefore confers sufficient free will to the agent and allows for an ascription of responsibility (on this occasion and for future acts which might be determined by the freely formed CEP):

1. Successfully choosing one or other option, despite a probability of failing to choose it and doing something else – despite it being undetermined which will be chosen.
2. Successfully doing what she *wanted* to do (even if what she "wanted to do" encompassed more than one course of action).
3. Psychologically/phenomenologically endorsing the choice as *hers* and recognising the choice as something she wanted to occur.

By this rationale, so long as one of Jessica's main competing desires wins out – and she therefore succeeds in doing something she wanted to do – and she then consciously endorses the ultimate decision as indeed what she wanted to do, we can say her eventual action was free in the Gold Standard libertarian sense. The key question we need to ask here is: does Kane's imposing of such conditions make his SFAs any less susceptible to the Luck Objection?

To answer this question, we need to consider whether these conditions address the fundamental point the Luck Objection is making, which is that if the agent could do or do otherwise *all past circumstances being exactly the same* – which includes all the agent's mental events – it seems there is nothing *within the agent* that can be deciding the factor in the decision and thus the ultimate decision can be argued to be a matter of luck. Considering Kane's account, we could very well grant that if Jessica ignores the fallen male and hurries on to her race, she would have a succeeded in choosing a course of action she desired and, were she to be questioned after, would acknowledge that going on to the race was the culmination of *her* internal deliberation and the *doing* of something *she wanted to do.* But none of that addresses the important issue of there being an effectively equal and opposite desire which she did *not* choose and there being no apparent explanation as to *why*. Jessica could meet all of Kane's conditions for a free SFA yet, if we were to wind back the world, she could meet them all again but take the opposing path. With nothing in her CEP to tip the balance one way or the other should we not conclude the eventual outcome is indeed just a matter of luck?

Kane wishes us to accept that the term “luck” is effectively being *misused* in such discussions and that in such cases as his SFAs, where all his conditions are met, the situation may still be undetermined but does not deserve to be saddled with the connotations that the familiar usage of the term carries (and the way the proponents of the Luck Objection intend the term to be taken), and thus does not interfere with the agent acting freely. Even if we were to accept this semantic argument and move on from using terms such as “luck” or “chance,” we are left with questions over the extent of the agent’s *control*. This aspect of Kane’s approach will be discussed in greater detail later on in this section, but the issues are entwined and our views on Kane’s success in establishing sufficient control will likely come down to whether we consider the lack of any identifiable “tipping point” from within the agent (which swings the balance from one course of action to another) as an important factor in the course of human action and decision-making. Kane’s view is that the only way such a tipping factor could exist would be in a world of antecedently determining control, which is simply not available on a libertarian view (at least not for the key SFAs) and which, if universal, would actually preclude free will. Thus, such a tipping factor is not required.

Whichever side of this debate you find more compelling, it could be argued that there does seem to be something important *missing* from Kane’s model: an explanation, based entirely on the agent’s CEP, for *why* one outcome was chosen over another; rather than just the passing of some arbitrary mental threshold. Perhaps the Luck Objection – in answer to Kane – could be renamed “The Control Objection” or “The Arbitrariness Objection” but it remains a thorn in the libertarians’ side as it raises a potentially required condition they seem unable to meet.

2.3.5 The Basic Argument

Whereas the Luck Objection focusses on the argument that there is an inevitable (and responsibility reducing) arbitrariness which results from the apparent break in causal influence generated by compliance with the Indeterminist (or AP) Condition, the Basic Argument attacks from the other side of the *Catch-22 of Libertarianism*; arguing that the *lack* of any break in causal influence would inevitably lead to a vicious regress, whatever the truth concerning determinism or indeterminism.

As discussed in section 1.2, the Basic Argument is designed to show that any attempt to ascribe ultimate responsibility to an agent can lead only to an infinite regress as it is not possible for an agent to be *causa* *sui* (the cause of herself). This means that even if the universe should prove fundamentally indeterministic free will is still impossible because a decision or action carried out at any point in time for which we are to be held responsible must flow from a character (or CEP) inevitably formed by antecedent events; antecedent events which flowed from a character at a previous instant that was formed as a result of previous decisions and actions…and so on.

Let us remind ourselves of the argument's structure, as based on the work of Galen Strawson:

1. We act as we do, in the present, due to our current established psychology (CEP) – the way we are now, mentally speaking – as developed by our heredity and our experiences.
2. If we are to be responsible for how we act in the present, we must be ultimately responsible for the formation of our CEP.
3. But we cannot be ultimately responsible for the formation of our CEP because, in order to be so, we would have to have carried out some actions in the past, for which we can be held responsible, which contribute to the formation of our CEP.
4. And in order to be responsible for those actions we must have been responsible for the formation of our established psychology (EP) that was in operation at that time in the past.
5. But, in order to be responsible for the formation of our EP that was in operation at that time in the past, we would need to have been acting responsibly at an earlier time still, which would contribute to the formation of this past EP.
6. And so on, back to a time in our lives when we could not be said to be capable of choosing anything.
7. So free will is impossible because, ultimately, we cannot be the cause of ourselves.

Kane's response to this argument is to broadly accept it for many (and probably most) of our actions but to argue that his SFAs are the exception as they are not in any way causally necessitated by an agent's antecedent psychology. Kane is therefore positioning his SFAs as "regress-stopping" because they do not have sufficient causes or motives for occurring; they are not *determined* by an agent's CEP. Rather (as discussed above), such desires are born out of the indeterminism-creating inner turmoil of a torn psyche in mental conflict. The desires remain based in the agent's CEP but, as there is more than one in genuine competition, parallel streams are created within which the competing desires battle it out for supremacy. The winning desire is something the agent wanted to achieve all along and thus Kane would argue that the decision can be said to flow from her character, but not sufficiently for the outcome to be determined prior to it coming to be. Responsibility for Kane, therefore, does not rely on sufficient causation. So long as his own conditions for a successful SFA are met (choosing despite probability of failure, doing something you wanted to do, endorsing the decision as your own, and so on) then the act is free and responsibility is conferred automatically.

Kane's alternate structure might thus proceed as follows:

1. We [mostly] act as we do, in the present, due to our current established psychology (CEP) – the way we are now, mentally speaking.
2. If we are to be responsible for how we act in the present, we must be ultimately responsible for the formation of our CEP.
3. We *can* be ultimately responsible for the formation of our CEP, but only if *some* of the actions we have carried out in the past were the product of undetermined willings (SFAs) which meet the criteria for which we can be held responsible for them.
4. Such actions have no sufficient causes or motives, and thus the regress is halted.

It might seem as though – by positing his regress-stopping SFAs – Kane is arguing that (*contra* Strawson) an agent such as Jessica *can* in fact be *causa sui*, the cause of herself, but Kane may not wish to go quite that far. During SFAs, when Jessica's self-network is working on its two incommensurable options and generating indeterminism deep in the neurons of her brain, the past is "screened off" (as Kane puts it), the future is undetermined and she is actively trying to achieve two different courses of action, one of which she will succeed in doing. The point, I contend, Kane would wish to emphasise is that the course of action ultimately chosen will have been born out of her CEP, *whichever way she goes*. Thus, he is not presenting an image of a Godly agent stepping entirely outside of the causal order and manufacturing new physical causal chains out of a void. What Kane is actually describing is a process of linear, but *branching*, causation which an agent can take advantage of to allow for different outcomes to indeterministically fight it out.

The problem for Kane is that in acknowledging that Jessica is *not* causing herself as some *prime mover unmoved,* he potentially opens himself up to the charge that Jessica is not, in fact, causing *anything at all*. Having initiated the branching causal process, Jessica appears powerless to decide the outcome of the ensuing competition. Either the desire to stop and help will win out or the desire to drive on will, and there is nothing to explain *why* one desire wins out over the other. The eventual SFA, therefore, that modifies her CEP and adds to the creation of her character going forward does not appear to have been caused by the agent, but is rather a *random* result of her branching mental cognition and ensuing indeterminate competition. And if, when we act in the present, we are acting due to a CEP formed by a combination of heredity, experience, *and luck* (or at least *randomness*), we seem no better off than when we were subject to the regress of it being formed by heredity and experience alone. In this way, for Kane's theory, the problems posed by the Luck Objection and the Basic Argument can be reduced to the same overriding issue: there appears to be nothing at all *within the agent and her deliberating process* that can explain why one course of action ends up being chosen over another at the culmination of the parallel processing, so the eventual outcome appears solely down to randomness and cannot therefore be one for which she can be held responsible.[20]

There is a further issue that – even if we were to accept everything Kane is arguing – there is a case to be made that the regress is not actually halted entirely. In the case of Jessica, for example, we might allow that the outcome of the choice is undetermined but could we also comfortably allow that the options which arise for consideration are also entirely undetermined? Such options would presumably flow forth determinedly from her CEP, but then it appears this would just lead to a different form of infinite regress; one where we can only be responsible for the options which arise for consideration as part of a SFA (which in a significant way control the outcome) if we were also responsible for the previous options that arose for consideration during the previous SFA…and so on, right back to a time before our first SFA. On seeing the man fall, Jessica's options in our example are to get out and help or drive on but it is conceivable that, were she to be a very different person (with perhaps significant trauma in her past), her only truly considered available options could be to drive on or even turn the wheel and drive over him. Thus, the arising options could be argued to play just as significant a part in the formation of an agent's CEP as the decision itself.

[20] This indeed appears to be Strawson's view: "In Kane's view, a person's 'ultimate responsibility' for the outcome of an effort of will depends essentially on the partly indeterministic nature of the outcome. This is because it is only the element of indeterminism that prevents prior character and motives from fully explaining the outcome of the effort of will. But how can this indeterminism help with moral responsibility? How can the fact that my effort of will is indeterministic in such a way that its outcome is indeterminate make me truly responsible for it, or even help to make me truly responsible for it?" (Strawson, G. 1994, p.21)

Mele makes a similar point by focussing on the *trying* to make the effort, one stage up from the *making* of the effort. In doing this he seeks to highlight a disanalogy between the assassin case and Jessica's situation that shows the supportive move from the former to the latter in Kane's argument is illegitimate.

> In Kane (1999b), it is not claimed that in cases of dual efforts to choose, the choices made are products of freely made efforts. Nor did Kane put himself in a position to make this claim, for that article includes no account of what it is for an effort to choose to *A* to be freely made. Thus, there is a striking disanalogy between cases like that of Kane's assassin and Kane's dual trying cases: no grounds are offered for presuming that the dual efforts to choose are freely made. And if the agent's efforts to choose in a dual trying scenario – unlike the assassin's effort to kill the prime minister – are not freely made, it is hard to see why the choice in which such an effort culminates should be deemed free. (Mele in Palmer (ed.) 2014, p.39)

Mele then goes on to refer to the already discussed example of the neuroscientist manipulating an agent into wanting what he (the neuroscientist) wants him to want and therefore controls what the agent tries to bring about. Mele argues that the tryings may be internally indeterministic but the agent does not freely try to make the choices he tries to make. Mele concludes that if we take away the puppet master, the puppet's position has not improved. According to Mele, then, Kane's "regress stopping" SFAs are not really regress stopping at all because if the efforts resulting in a choice are themselves to be free then we need to postulate further SFAs in order to initiate the efforts and further SFAs to initiate the initiating of the effort, and so on.

Mele's specific objection fails, however, because why should it matter if the initial requirement to act at all – and thus the instigation of an effort – is guided initially by circumstance? This fact in itself surely does not preclude freedom and responsibility if the response to such a circumstance is the process of an SFA. It does not seem necessary for us to somehow freely choose to respond each and every time something occurs in front of us that would likely occasion a response. It is enough that such efforts often, if not always, arise organically. Kane, it seems, agrees with this particular point:

> The plural efforts preceding SFAs *might* have been initiated by further SFAs in certain cases. Agents may sometimes be conflicted about whether even to *begin* to *deliberate* about a difficult choice they have an aversion to thinking about. But this need not always be the case and often will not be the

> case. The plural efforts preceding self-forming choices will normally be initiated by the confluence of the agent's conflicted will *plus* the agent's recognition of the situation he or she is in. (Kane in Palmer (ed.) 2014, p.199)

However, this response does not answer *my* objection regarding the *content of the options themselves* and whether their arising organically on each occasion (either deterministically or non-deterministically) undermines Kane's succeeding SFAs, and thus his whole model for libertarian action.

As with the Luck Objection, I do not consider Kane's response to the Basic Argument to be conclusive. Kane could (and does) argue that the Basic Argument's apparent assumption that responsibility is predicated on some form of sufficient causation by the agent's CEP is begging the question against an indeterminist account but, as ever, the difficulty is in presenting an acceptable libertarian alternative. Again, as with the Luck Objection, our view of whether Kane's responses are in any way successful therefore comes down to whether we consider Kane's conditions for a successful SFA as an acceptable alternative to the potentially more familiar (and ostensibly simpler) model of antecedently sufficient causation. At this point I could not say with complete confidence that they are.

2.3.6 Chance, Responsibility and Control

Because he is forced to acknowledge the difficulties that objections such as the Luck Objection and Basic Argument present for his view, and the near impossibility of answering them to everyone's satisfaction, Kane seeks out varied (and occasionally more speculative) avenues of support. To restate the main issue he is facing, his theory (and indeed, all libertarian, event-causal theories) generates the following, seemingly unavoidable, consequence: there appears to be nothing that can explain, in terms of an agent's CEP, why one course of action (or desire or decision) wins out over any other in dilemma cases such as the one faced by Jessica and so what course of action is ultimately followed, it has been argued, must only be down to luck. To put it another way: if nothing in the description of the world at time t (including everything there is to know about an agent) entails whether Jessica will help or not, and she could therefore just as easily do either, it is difficult to see how the ultimate choice could be in her *control* in the sense required for any ascription of responsibility.[21]

Even if we were to accept the whole of Kane's argument – accept that, on witnessing the man fall, Jessica is compelled toward two incommensurable behavioural options, which stirs up chaos in her brain and generates physical indeterminism at the quantum level in her neurons, which is magnified to the level of conscious deliberation, splitting her psyche in two, leaving two versions of herself, each attempting (one-way rationally) to achieve a contrasting objective – even if we accept all of that, there is simply no explanation for why one Jessica wins out over the other. If the selfish Jessica wins out and she drives on to her race meeting, we are left without any answer as to what caused her selfish desires to overcome her moral ones which, as we have seen, leads some to argue that libertarian approaches such as Kane's depict an unacceptable reality in which choices can seemingly be a product of chance and yet we can still be held responsible for them.

As briefly discussed above, Kane attempts to address the issue partly by focusing on semantic considerations – particularly those surrounding the terms "luck," "chance" and "control." He acknowledges that critics will perhaps consider his entire theory, and his responses to the criticism of it, and still claim that if Jessica and Jessica* have exactly the same past, prior motives, character – and thus an identical CEP – and both make the same efforts to both do the selfish thing and do the non-selfish thing right up to the moment of decision and yet ultimately make different choices, then some form of chance must be involved. But, Kane argues, even if this is true, where an agent is trying to achieve two mutually exclusive outcomes, they cannot but help be responsible for which outcome they choose, whichever way it goes. This is because he believes the chance involved is not the same chance as that of a coin flip or a dice roll, and an element of – indeterministically inspired – chance being involved is not the same as saying "they chose *by chance*."

[21] John Martin Fischer puts the objection as follows: "Let us suppose that causal indeterminism (of this sort) obtains, and that I choose at time *t*2 to raise my hand at time *t*3. It follows that a statement of the total set of facts that obtained at *t*1, together with a statement of the laws of nature, fails to entail that I choose at *t*2 to raise my hand at *t*3. Thus, everything that obtained just prior to *t*2, including everything about me – all my preferences, beliefs, dispositions, traits of character, and so forth—is completely compatible with my *not* choosing at *t*2 to raise my hand at *t*3 (even on the supposition that the laws of nature are held fixed). Given this implication of causal indeterminism (of the sort under consideration), it can seem puzzling how my actual choice at *t*2 to raise my hand at *t*3 could really be *mine* – could be in my control in the sense required for moral responsibility. After all, everything about me could be the same and yet I *not* make this choice." (Fischer in Palmer (ed.) 2014, p.53)

> To say something was done "by chance" usually means (as in the assassin and husband cases when they fail), it was done "by mistake" or "accidentally," "inadvertently," "involuntarily," or "as an unintended fluke." But none of these things holds of the businesswoman and John either way they choose. Unlike husband*, businesswoman* and John do not fail to overcome temptation by mistake or accident, inadvertently or involuntarily. They consciously and willingly fail to overcome temptation by consciously and willingly choosing to act in selfish or weak-willed ways. So, just as it would have been a poor excuse for the husband to say to his wife when the table broke that "Luck or chance did it, not me," it would be a poor excuse for businesswoman* and John to say "Luck or chance did it, not me" when they failed to help the assault victim or failed to arrive on time.[22] (Kane 1999, p.235)

By making the implicit connection between indeterminism and the ordinary language terms of "luck" and "chance," argues Kane, his critics are begging the question against indeterminist accounts and, in reality, the implications of the latter terms are not necessarily also the implications of the former. Indeterminism for Kane (as discussed above) implies only the absence of deterministic causation not necessarily any of the familiar connotations ascribed to the terms "chance" or "luck" as well. Saying the Jessica who stopped to help "got lucky" can mean succeeding in the face of a probability of failure, but this is not the same as saying the outcome was not of her doing or occurred by mere chance; and it certainly does not mean she was not responsible for it. Moreover, Kane argues, agents such as Jessica are clearly exercising voluntary control over both potential outcomes when considered as a set – what he terms *plural voluntary control* – despite the evident diminished control over each option when considered separately, due to the indeterminism involved being created by the conflict in her will. He is indeed acknowledging that indeterminism does diminish control and hinders or obstructs our purposes but, in the case of SFAs, that indeterminism comes from *within our own will*, from the desire to do opposite, and so does not remove the responsibility for our actions.

Importantly, Kane fully accepts that his plural voluntary control is not the same as *antecedently determining control*, which many (such as compatibilists, hard determinists and some libertarians)[23]

[22] The examples referenced are of the assassin aiming at the PM, the businesswoman who witnesses a robbery, John who needs to be on time for a meeting (Mele), and a husband who does (or does not) break a glass table.

[23] Libertarians who subscribe to the agent-causal approach generally see the causation enacted by an agent as determining, though this causal activity of the agent is itself uncaused and thus undetermined. More on this in sections 2.6 and 2.7.

believe is the only kind of control which could possibly be sufficient for moral responsibility. But, again, antecedently determining control, according to Kane, is not the only kind of control a person can have. Indeed, he maintains that those who believe it is the only kind of control a person can have will always be left wanting when trying to achieve true (libertarian – Gold Standard) freedom of the will as this kind of control can only ever predetermine the outcome in a fashion which would be unacceptable to libertarians. Kane is therefore proposing a different kind of control:

> On the account given, however, the agent does have control over the occurrence or nonoccurrence of the choice made in the following important sense: to have control at a time in this sense over the being or nonbeing (the occurrence or nonoccurrence) of an event (e.g., a choice) is to have the power, ability, and opportunity at the time to make the event *be* at the time and the power, ability, and opportunity at the time to make the event *not be* at the time. The agent not only has this power at the time to make either choice be, but also the power to make it not be, *by making the competing choice be*. The agent has both these powers because either of the efforts in which the agent is engaged might succeed in attaining its goal despite the chance of failure. (Kane in Palmer (ed.) 2014, p.195)

The difficulty for Kane, is that it seems open to critics to simply disagree and argue that the agent just does *not* have the power to make "either event be at the time." Rather, she has only the power to set in motion a deliberative process (and even this may only be able to arise deterministically) which will eventually culminate in a choice one way or the other, with no further interventions by the agent. It could also conceivably be claimed that Kane is arguing by stipulation[24] as he is defining "control" in such a manner as to fit his theory without perhaps offering sufficient justification for why his definition should be accepted over available alternatives.

Perhaps the clearest signal that Kane is aware his position leaves certain key questions unanswered or key issues unresolved, is his inclusion of more speculative ruminations on accounting for the arbitrariness that seems inherent in his model.

> [T]he question of arbitrariness is one that must be forthrightly addressed. I grant that free choices are arbitrary in the sense that they are not fully explicable in terms of the past. But that is because

[24] Something he has accused other proponents of competing libertarian theories of doing. See sections 2.5 and 2.7 for examples of this. See also Kane 2005, p.48, 51 and 58.

> their full meaning or significance does not lie in the past…Every [SFA] is the initiation of a "value experiment" whose justification lies in the future and is not fully explained by the past. It says, in effect, "Let's try this. It is not required by my past, but it is consistent with my past and is one branching pathway my life could now meaningfully take. I'm willing to take responsibility for it one way or the other. It may make me a (morally) better or worse person, a (prudentially) happier or sadder person, but I will be aware that it was my doing and my own self-making either way – guided by my past, but not determined by it. (Kane 1996, p.145)

And, going even further:

> It is worth noting in this connection that the term "arbitrary" comes from the Latin *arbitrium*, which *means* "judgment" as in *liberum arbitrium voluntatis* or "free judgment of the will" (the medieval designation for free will). I have argued elsewhere that there is a kernel of insight in this etymological connection for understanding free will…free agents are both authors of, and characters in, their own stories all at once. By virtue of their free choices (*arbitria voluntatis*), they are "arbiters" of their own lives" – “making themselves" as they go along out of a past that (if they are truly free) does not limit their future pathways to one. (Kane 1996, p.145)

It can be argued, though, that these semantic considerations, while interesting, do not address the key issue his critics are raising, which is the lack of any deciding factor (internal to the agent) when an agent is faced with two competing alternatives. If there is no explanation for why an agent making efforts to realize such competing objectives succeeds in realizing one over the other, the outcome arguably appears to be the result of random processes and, whilst Kane may be correct that a lack of antecedently determining control does not necessarily entail *chance* or *luck*, this does not, of itself, mean that the kind of control he is proposing (were it to be accepted) would be sufficient for many people’s conception of genuine responsibility. Kane, of course, is free to argue to the contrary, but his argument is unlikely to win over anyone from an opposing camp. The fact remains that if indeterminism means that Jessica and Jessica* (in different possible worlds) can be exactly the same in all relevant ways right up to the point of decision and yet still remain able to choose to follow completely different courses of action, it will always be open to critics to argue that libertarians must accept a level of uncertainty in nature which raises difficult questions for undetermined free agency.

2.3.7 The *Criteria*

Robert Kane has proposed a detailed and sophisticated account of libertarian free will, which I will now assess against my *Criteria for Coherence* as a final judgement on its intelligibility. Let us begin with the first three principles taken together:

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.
P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.
P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.

Kane's event-causal approach is clearly an incompatibilist theory resting on a platform of indeterminism and involves agential decisions that are not fully determined by past conditions; thus, adequately meeting the first two principles. In general, his model would seem also to meet the third principle with much talk of his agent requiring the ability to do and do otherwise in order to be free and responsible. It is worth noting, however, that there are times in his works where in his more speculative musings his adherence to this point no longer becomes quite so clear. Consider the following quote discussing indeterminate (non-determined) efforts, where he is trying to counter the claims of the Luck Objection:

> With indeterminate efforts, exact sameness is not defined. Nor is exact difference either. If the efforts are indeterminate, one cannot say the efforts had exactly the same strength, or that one was exactly greater or less great than the other. That is what indeterminacy amounts to. So one cannot say of two agents that they had exactly the same pasts and made exactly the same efforts and one got lucky while the other did not. Nor can one imagine the same agent in two possible worlds with exactly the same pasts making exactly the same effort and getting lucky in one world and not the other. Exact sameness (or difference) of possible worlds is not defined if the possible worlds contain indeterminate efforts or indeterminate events of any kinds. And there would be no such thing as two agents having exactly the same *life histories* if their life histories contain indeterminate efforts and free choices. (Kane 1996, p.171)

Kane appears to be claiming here that – despite his central arguments that two versions of an agent could have exactly the same past *and* still be responsible for the competing outcomes of an SFA – it may in fact not be metaphysically possible for Jessica and Jessica* to make exactly the same efforts in the run up to their decision, due to fact that the effort itself is indeterminate (not determined by past conditions). This could well be correct, but would seem to contradict (or at least invalidate) his own Indeterminist Condition (which states that agents must be able to act or act otherwise given the exact same past circumstances and laws of nature) as it would mean duplicate agents in different possible worlds could never have the same past circumstances at all. This rather sharp turn (which seems added almost as an afterthought) could be taken to further indicate that Kane is not actually truly comfortable with the concept that an agent can act or act otherwise given exactly the same past circumstances (or even as a tacit admission that such a concept is generally incoherent) and puts his conformity with principle three into doubt.

P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.

Given his semantic considerations regarding the common usage of terms such as “arbitrary,” “luck” and “chance,” Kane would no doubt claim that he has met this principle, but it is not clear that his arguments – that whilst there may inevitably be some form of arbitrariness in the libertarian decision-making process, this would nonetheless *not* diminish agential control (and thus responsibility) – would be readily accepted by any proponents of the Luck Objection. Additionally, his allowance that there can be ubiquitous cases of determined decision making (non-SFAs flowing from a freely formed will) could lead some to argue that his theory actually offers a model for decision-making that is *both* arbitrary (in the sense of SFAs) *and* determined (in the sense of the non-SFAs). Whilst it is perhaps not completely clear whether Kane could be said to have met this principle, there are good arguments to say that he has not.

P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.

Kane has certainly attempted to address this issue head on with the development of his SFAs and his presentation of them as regress-stopping but, as we have seen, some form of regress seems unavoidable on this view due to the determined nature in which the *options for consideration* must arise. In any case, even if we were to concede that his SFAs *are* truly regress-stopping, the nature of *how* they halt the regress is what leaves his theory open to the charge that it would entail an unacceptable level of control-reducing arbitrariness in the decision-making process. Either way, it can likely be argued that this criterion has not been met.

P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

With the introduction of his Free Agency Principle, it would seem his success in this regard was guaranteed. However, the reliance on phenomenology and conscious experience in order for an agent to essentially validate their choices as their own could be seen by some as the introduction (or at least the unnecessary invocation) of an extra mystery competing theories do not necessarily need to avail themselves of due to the lack of widespread agreement on the nature and structure of both those mental aspects. Also, some might argue that his focus on the re-defining of familiar terms and his speculation concerning "future justification" for present decisions and "value experiments" is somewhat vague and opaque. Still, on balance, I believe it would be fair to argue that this criterion has been met.

2.3.8 Conclusion

In summary, despite promising developments concerning the ascription of (some form of) control and perhaps even (a restricted form of) responsibility, Kane's theory has not ultimately fared well against my *Criteria for Coherence* (and thus against his own stated goals), failing to meet its key principles. I do not believe it can therefore be considered a fully coherent theory of free will. What Kane's efforts *have* shown is just how difficult it is for libertarians to escape from the *Catch-22 of Libertarianism* that they have seemingly theorised their way into. The question for the rest of this thesis then is: can we improve upon Kane's SFAs and thus present a model for decision-making that can be both regress-stopping and also entirely non-arbitrary?

2.4 Non-Causal Theories

Non-causal (or simple-indeterminist) approaches attempt to overcome the problems posed by both deterministic and indeterministic causation by arguing that free actions need not necessarily be caused at all by an agent's antecedent states; their desires or beliefs and so forth. They can be entirely uncaused, they claim, with no internal causal structure to speak of, and yet still be considered non-random, under the control of the agent and can be explained in terms of agent's reasons and purposes.

Doing away with any causal link completely, however, raises some immediate and obvious problems that such theorists need to address. Firstly: if an action was not caused by the agent (or states within the agent), how can we say that it was *her* action, brought about by her in a manner sufficient for her to be considered responsible for it? How did she determine that *that* particular action and not some other alternative one took place? In short: how was the action in her control? Secondly, without any causal link between an agent's antecedent states and her action how are we to explain that action in terms of her *reasons for doing it*? Typically, we expect an agent to have reasons for acting in a certain way. If it is not possible to say that having/recognising those reasons *caused* the action (as on the event-causal approach) then in what other way can we provide a suitable reasons explanation?

Non-causalists argue that to assume a causal explanation is what is required is simply begging the question in favour of a causal account and claim that free actions can be uncaused and still occur for reasons in a satisfactorily non-arbitrary fashion. An agent's reasons, they argue, can influence her actions without necessarily causing them and do so, typically, by entering into the contents of her intentions.

2.5 Carl Ginet's Sufficient Conditions Approach

One of the most prominent defenders of the non-causal approach is Carl Ginet, who argues that the way an agent can have active control over her actions and be able to provide an explanation for such actions in terms of her reasons – without positing a causal link between her antecedent states and her action – is simply by the act of *performing* the action and by being the *subject* of the action.

> …any attempt to explain an agent's determining of her own action in terms of this or that special sort of cause of it is unnecessary. For an agent to determine whether or not an event, *e*, occurs is for her to make it the case that *e* occurs by performing some suitable free action. If *e* is *not* her own free action then causation must enter into what it is for her to make it the case that *e* occurs: she can do this only by performing some free action that causes *e*. But if *e is* her own free action, then she makes it the case that *e* occurs, not by causing it, but by simply performing it…Given that the action is free, the agent determines it, one could say, simply by being its subject, the one whose action it is. That is to say, all free actions are *ipso facto* determined by their subjects. (Ginet 1997, p.87)

Actions, Ginet argues, start with a basic, causally simple, mental event not caused by any antecedent mental events and what makes such mental events an action at all (as opposed to an event that merely happens) is an accompanying *phenomenal* quality, which Ginet calls the "actish" quality. It is this phenomenal quality that makes the agent the *subject* of her action. It is what makes it seem to the agent that it is *her* action and not just something that happens to her. What then makes such an action *free*, is the fact it is not causally necessitated by antecedent events.

> Every action, according to me, either is or begins with a causally simple mental action, that is, a mental event that does not consist of one mental event causing others. A simple mental event is an action if and only if it has a certain intrinsic phenomenal quality, which I've dubbed the "actish" quality and tried to describe by using agent-causation talk radically qualified by "as if": the simple mental event of my volition to exert force with a part of my body phenomenally seems to me to be intrinsically an event that does not just happen to me, that does not occur unbidden, but it is, rather, as if I make it occur, as if I determine that it will happen just when and as it does (likewise for simple mental acts that are not volitions, such as my mentally saying "Shucks!").
> (Ibid., p.89)

Ginet explains that wider biological activity can indeed be causally complex, as many would naturally expect it to be, but the initial, simple mental action must be causally simple and thus not caused (either deterministically or probabilistically) by other events, such as an agent's desires or beliefs. In order for an action to be free and responsible in Ginet's sense, then, it must conform to two main criteria. It must not be causally necessitated in any sense, which makes it *free*, and it must be accompanied by this actish phenomenal quality, which confers sufficient control (and thus responsibility) to the agent whose action it is.

As stated above, this approach raises a number of questions which need to be addressed. What, ontologically speaking, *is* this initial, causally simple, mental event that Ginet is proposing? How does it arise and how is it linked to an agent's current established psychology (CEP)? Can a phenomenal experience of volition, by itself, metaphysically confer sufficient ownership of – and control over – an action to the agent? How can we provide a satisfactory explanation in terms of an agent's reasons or purposes in the absence of a causal connection between an agent's desires and her actions? If the explanatory connection is *not* causal, then what is it? Such questions will be considered in what follows.

2.5.1 Control Without Causation?

The first issue to consider is Ginet's account of active control, which seems to be based wholly on the existence of a phenomenal quality, which is intrinsic to mental actions of the appropriate type. It is not initially clear what a "simple mental event" of the sort Ginet is describing actually is, nor is it clear how its having this "actish" phenomenal component would confer control upon an agent in such a manner as to position Ginet's model as a credible alternative account to causal explanations. Even were we to accept that a direct causal link between an agent's antecedent mental states (desires, beliefs, deliberations and so forth) and her eventual active volition was not essential for active control, it does not initially seem enough to have this actish phenomenal quality alone or seem easy to accept that an agent could determine her action simply by being the *subject* of that action, as Ginet argues. The approach seems instinctively counter intuitive and thus requiring of further explanation.

In his main text on the subject, *On Action,* Ginet explains that the simple mental action that comprises an agent's volition is the start of any voluntary exertion (such as Jessica's opening of the car door and getting out to help the elderly man). Such acts, he claims, are an "exception to the claim that acting consists in causing something” (1990, p.11) and as they are the root of a new string of action, with no causal antecedents, they can therefore be considered as regress-stopping – a key goal for libertarian theories in the attempt to meet the UR condition (as we have seen in previous sections). The bulk of his explication of such basic acts, however, does not focus on acts such as this, rather it is taken up with the consideration of a different sort of mental event to volition: that of mentally saying a word. Such a mental occurrence, he claims, does not have the structure of one event causing another, as it is clearly not two distinct causally related events with the first causing the second. Nor, though, is it like the unbidden occurrence of a word in the mind of an agent where the agent does not mentally *say* it but just hears it (perhaps out of the blue), as such an occurrence, he argues, would not count as an *act*. In order for it to count as an act on Ginet’s model, it requires his actish phenomenal quality. Ginet gives the following description and claims that it applies also to the key simple mental act of volition (or willing):

> The unbidden occurrence is not an act…Rather, the mental act differs from the passive mental occurrence *intrinsically.* The mental act has what we may call (for lack of a better term) an *actish* phenomenal quality. This is an extremely familiar quality, recognizable in all mental action, whether it be mentally saying, mentally forming an image, or willing to exert force with a part of one's body. The only way I can think of to describe this phenomenal quality is to say such things as "It is as if I directly produce the sound in my 'mind's ear' (or the image in my 'mind's eye', or the volition to exert), as if I directly make it occur, as if I directly determine it" - that is, to use agent-causation talk radically qualified by "as if". This quality is intrinsic to and inseparable from the occurrence of the word in my mind when I mentally say it. It belongs to the manner in which the word occurs in my mind and is not a distinct phenomenon that precedes or accompanies the occurrence of the word. Similarly for the act of forming a mental image. (Ginet 1990, p.13).

A number of objections have been raised against this account of active control. One objection, discussed by Randolph Clarke (as well as Ginet himself), is that his actish phenomenal quality could just be illusory and thus not convey any *real* control at all. Ginet in the quote above uses the words “as if” as in: “*as if* I directly determine it” (ibid, p.13, my italics); indeed, apparently leaving open the possibility that it only *seems* as if we are in control of our actions when, in reality, we are not.

As Clarke explains:

> Whatever the correct characterization of this phenomenal quality, the mere feel of a mental event – the way it seems to the individual undergoing it – although it may be a (more or less reliable) sign of active control, cannot itself *constitute* the agent's exercise of such control…To hold that it does is to render the exercise of active control wholly subjective (nothing more than the way things seem), and this is to greatly diminish the significance of active control. (Clarke 2003, p.20)

If we are not really in control – if our actions are not determined to occur *by us* – and the actish quality is thus only illusory, volitions would then seem to fall into the same category as the unbidden mental occurrences Ginet stated were *not* actions at all. Volitions arising in this manner, without any identifiable cause and no other apparent substantive link to the agent's CEP, appear to take us back to the possibility of our nefarious neuroscientist – discussed in section 2.3 – who sets about implanting volitions into the mind of an agent, which then take the form of the initial uncaused mental act that ultimately results in this or that behaviour. Such implanted volitions would still have the intrinsic actish phenomenal quality – and thus be indistinguishable from volitions *not* implanted by a malign third party – but, surely, we would not wish to say that such a volition and its subsequent action were fully under the control of the agent herself. This, it could be argued, would be another example of an agent being capable of free action (in Kane's one-way voluntary sense) whilst not fully having the Gold Standard freedom of the will. The argument is, therefore, that the actish phenomenal quality does not seem enough in and of itself to make a mental event a basic action of the type sufficient to confer active control. Whilst it could be argued that this is begging the question against Ginet's account – and that he is arguing that "control" in his sense, is simply not a causal notion – the question remains whether the phenomenal "link" is enough to entail such a notion as responsibility, especially if such a link could be seen as only illusory. Ginet himself concedes that some will indeed have such concerns, but then really only seeks to counter them by arguing that the causal alternatives to his theory (agent-causation in particular, which I shall deal with in detail in section 2.7) do not provide successful accounts of free will and themselves would also fall prey to a many of the objections levelled at his theory.

Additionally, concerning the basic actions themselves and their lack of causal properties, Ginet argues that this is so, *by definition*: they just do not have a sufficiently complex causal structure to enable them to be caused by anything which, if true, would seem to invalidate most other libertarian positions based on event or agent-causation.[25]

> The simple mental act presents much the same problem for the thesis that an action is always the subject's *agent*-causing something as it does for the thesis that it is always the subject's *event*-causing something. The simple mental act simply fails to have a sufficiently complex structure. The actish quality of the mental occurrence is enough in itself to make the occurrence a mental *act*. A mental occurrence with that intrinsic quality is ipso facto a mental act. No *extrinsic* relation of that occurrence to its subject or to another event, no relation that the occurrence could have failed to have, is needed to make it a mental act. The simple mental act cannot consist in the subject's either agent-causing or event-causing some event, no matter what relation between the person and that event the agent-causal relation or the event-causal relation is supposed to be, as long as it is something more than merely the person's being the subject of the event. Since the agent- and event-causation analyses of what it is for a person to cause something are the only ones in the cards, simple mental acts are counterexamples to the thesis that acting is causing. (Ginet 1990, p.14)

Ultimately, Ginet (as others, including Kane, have argued)[26] appears to be arguing by stipulation; claiming that a mental act just *is* a mental occurrence with the actish quality and, if we accept this definition, it would necessarily entail the explanatory failure of causal theories. Yet, in simply assuming that basic actions are not sufficiently complex to be caused by anything, he appears to be begging the question in favour of his non-causal account – thus utilising the same tactic he accuses his detractors of employing in favour of their causal accounts. Ginet may well be correct about the shortcomings of the competing theories, but this in itself provides no reason to necessarily prefer his account. What is needed, if we are to take what would appear to be the counterintuitive step away from the standard causal model, is an understandable and viable alternative to how an agent can exert active control over her behaviour – which requires some form of link between her CEP and her decisions and ultimate action – which Ginet's theory does not at this stage appear to adequately provide.

[25] Ultimately, Ginet seems to fall back to a "that's just how it is" line of argument, regarding the actish quality conferring act/control status on a mental event: "In answer to the question "Why should an action's being undetermined and having a certain intrinsic property be enough to make it an event that the agent determines?" the agent-causationist, as much as the simple indeterminist, has to say, "Well, it just is"; and the property of agent-causation gives no better justification for saying this than does the actish phenomenal quality." (Ginet 1997, p.93)

[26] See Kane 2005, p.58.

2.5.2 Acting for No Reason?

The second line of attack Ginet needs to defend against is the view that non-causal theories are unable to satisfactorily explain an agent's actions in terms of her reasons for performing it. Typically, we assume the actions we perform are done so for reasons we hold and, by citing such reasons, we can provide a reasons-explanation of our actions. Causalist accounts, such as event-causation (as well as compatibilist accounts), base such explanations on the causal connection between the intentions or desires and the action, and this raises a number of important questions for non-causalist accounts such as Ginet's. Is the causal connection *required* to make reasons-explanations true? If we cannot say that reasons in any way caused the action of an agent, then in what other way can we offer an explanation in terms of those reasons? If an agent's recognition or acceptance of a reason or reasons for a decision does not cause the subsequent action, can that reason or reasons be considered a true reasons-explanation for the action?

Any view that claims a causal connection is not necessary must overcome Donald Davidson's famous argument (see Davidson 1963) that a causal connection *is*, in fact, required because an agent may well have a reason for carrying out a particular action, but just *having* that reason does not necessarily make it the reason *for* which the agent acts. Without the causal connection, what is it that can make a particular reason *the* reason for which an agent carries out a particular action? That is, what – in the absence of a causal connection – is sufficient for the truth of a reasons-explanation that references the reason in question? Ginet believes he has developed a satisfactory alternative by positing his "sufficient conditions" which, when met, he claims would guarantee the truth of a reasons-explanation.

Ginet's Sufficient Conditions

Ginet gives the following sufficient condition for the truth of a typical reasons-explanation (1) "S A-ed in order to B", which he labels (1-C): "Concurrently with her A-ing S intended of that A-ing that by it (and in virtue of its being an A-ing) she would B (or would contribute to her B-ing)" (in Kane ed. 2002, p.388). Ginet claims that, without entailing a causal connection between the intention and the action (or any instigating mental events), the truth of (1-C) *guarantees* the truth of (1) because one could not consistently affirm the former and deny the latter and therefore the proposition is sufficient for the truth of the reasons-explanation.

> My account of reasons explanations has it that when an action is correctly explained as one the agent did for a certain reason supplied by an antecedent motive (a desire or intention), the explanatory connection between the motive and the action is forged, not by causal laws (probabilistic or deterministic) or by agent-causation, but by an *intention* concurrent with the action. The intention has the following content: it refers directly to the current action and to the remembered prior motive and says that this action is to satisfy that motive (or to help to do so).
> (Ginet 1997, p.96)

There are two parts to Ginet's sufficient condition, then. Firstly, prior to the action, *A* has an antecedent desire or motive she wishes fulfilled. Secondly, concurrently with the action, *A* remembers the antecedent desire or motive and intends that by carrying out the action it will satisfy (or contribute to satisfying) the desire or motive. The antecedent desire or motive (the *reason* for the action) explains the action due to its being directly referred to by the concurrent intention and it is therefore not necessary that it also *cause* the action. The mere presence of the referential, concurrent intention is enough.

Returning to Davidson's challenge, where multiple reasons and thus differing concurrent intentions could be in play, it does seem such a situation would appear problematic for ascribing reasons-explanations where the explanatory connection is rooted only in the simple presence of a referential, concurrent intention. In such circumstances are Ginet's sufficient conditions really sufficient? Randolph Clarke has illustrated the problem as follows:[27]

> To illustrate the difficulty, suppose that an agent, Sarah, wants her glasses, which she has left in her friend Ralph's room, where he is now sleeping. Sarah also wants to wake Ralph, because she desires his company, but she knows that Ralph needs sleep right now, and hence she desires, too, not to wake him. Sarah decides to enter Ralph's room and does so, believing as she does that her action will contribute to the satisfaction of both the desire to get her glasses and the desire to wake Ralph. (The example is adapted from Ginet 1990: 145.) What further facts about the situation could make it the case that, in entering the room, Sarah is acting on the basis of the former desire, and that citing that desire provides a true reason-explanation of her action, while she is not acting on the basis of her desire to wake Ralph, and citing this latter desire does not give us a true reason-explanation of what she is doing?

[27] There is an interesting (and extensive) back and forth between Clarke and Ginet on this issue (See Ginet 2008, Clarke 2010 and Ginet 2016).

> Ginet's account of reason-explanations that cite antecedent desires (1990: 143) implies that the following conditions suffice for the truth of the explanation that cites Sarah's desire to get her glasses: (a) prior to entering the room, Sarah had a desire to get her glasses, and (b) concurrently with entering the room, Sarah remembers that prior desire and intends of her action that it satisfy (or contribute to satisfying) that desire. Given the indicated circumstances, citing Sarah's desire to wake Ralph will fail to give us a true reason-explanation, Ginet holds (145), just in case Sarah does not intend of her action that it satisfy (or contribute to satisfying) that desire.
>
> Suppose that, although conditions (a) and (b) are fulfilled as Sarah enters the room, her desire to get her glasses plays no role at all in bringing about (causing) her entry, while her desire to wake Ralph, of which she is fully aware when she acts, does play such a role. Is the content of her concurrent intention authoritative about what Sarah is aiming to do in entering the room, or is she fooling herself? Many will judge that, in this case, Sarah does not really act on the basis of her desire to get her glasses and that citing it does not truly explain her action. (Clarke 2003, p.22)

Clarke is making two related objections in this passage. Firstly, as Sarah has more than one desire and therefore potentially more than one concurrent intention – and none of these intentions are causally connected to her action – how can we, with any confidence, state which intention would feature in a satisfactory reasons-explanation for her action? How could we point to her desire to retrieve her glasses over and above her desire to wake Ralph and therefore cite that former desire as *the one* that explains the action? Furthermore, Clarke is arguing that despite holding the concurrent intention to retrieve her glasses, it could be the case that it is the second intention – the desire to wake Ralph – which brings about (or "causes") her to enter the room. If this were the case, then despite her desire to get her glasses being a concurrent intention (perhaps even to her mind the *main* concurrent intention), citing it would fail to explain her action.[28] Therefore, the mere *presence* of a concurrent intention is, in itself, not sufficient to explain an action.

[28] This type of causalist objection was developed previously by Mele, whose version of it – for an agent *S*, with two concurrent intentions *N* and *O* with regards to opening a window – goes as follows: "[S]uppose that a mad scientist, without altering the neural realization of *N* itself, renders that realization incapable of having any effect on *S*'s bodily movements…while allowing the neural realization of O to figure normally in the production of movements involved in *S*'s opening the window. Here, it seems clear, *O* helps to explain *S*'s opening the window, and *N* does not. Indeed, *N* seems entirely irrelevant to the performance of that action. And if that is right, Ginet is wrong; for on his view, the *mere presence* in the agent of an intention about her [action] is sufficient for that intention's being explanatory of her action." (Mele 1992, p.253)

In response to the second objection, Ginet accuses Clarke (as he does Mele)[29] of begging the question in favour of a causalist account and re-iterates his version of events: Sarah's act of entering the room was accompanied by her intention that by entering she would satisfy her desire to retrieve her glasses, which entails that she entered the room in order to satisfy that desire and so the desire was the reason (or one of the reasons) for which she entered.[30] To argue that a desire cannot be a reason for performing an action because it plays no role in causing the action, Ginet states, is simply question begging and he is correct in his assertion. Clarke (and Mele) are indeed just assuming a causal link is required to give a proper reasons-explanation and thus the lack of one is not sufficient. Clarke is attempting to conjure a scenario in which a concurrent intention can be present but explanatorily irrelevant and thus Ginet's conditions are not sufficient, but the explanatorily irrelevant concurrent intention is only so irrelevant due to the presence of a second concurrent intention stipulated to be causally efficacious. Removing the causal connection between the intention linked to the desire to wake Ralph and Sarah's action of entering the room leaves both concurrent intentions equally in play, explanatorily speaking, and restores Ginet's sufficient conditions approach as a contender. However, this then returns us to Clarke's first objection (*qua* Davidson): without the causal connection, what is it that can make a particular reason *the* reason for which an agent carries out a particular action?

Ultimately, in the face of this objection, Ginet appears just to re-assert his view and claim his conditions are indeed sufficient:

> For the example we have been using, the challenge is to supply something that will entail that a reason for which Sarah entered the room was her desire to get her glasses that does not entail that her desire caused (or played a role in causing) her action. But my account does that with its (as it seems to me intuitively compelling) claim that Sarah entered the room in order to get her glasses if concurrently with that action she intended of it that it contribute to satisfying her desire to get her glasses. So what is the problem? (Ginet 2008, p.236)

Davidson (and Clarke) are saying that an agent merely having a reason for an action does not make that reason the reason *for* which she acts and are asking what could it be, in the absence of a causal

[29] See Ginet, in Kane (ed.) 2005, p.389.

[30] See Ginet 2008.

connection, that *would* make it the reason for which she acts (and thus make that reason one that explains the action). Ginet answers that his condition of the agent having a concurrent intention to satisfy her desire is the extra thing that is needed and is thus sufficient for the truth of the corresponding reasons-explanation.

With those on the causalist side of the debate seemingly begging the question in favour of a causalist account and Ginet apparently trying to refute their objections by stipulation, the discussion – at least as framed here – would seem at an impasse with no clear way for one side to gain advantage over the other. It could be argued, however, that the debate in general rather misses the point of what is most important about debates on freedom and responsibility and that the emphasis on the problems of causation is directed to the wrong place in the action chain. It matters not whether a desire is causally efficacious in regards to an agent's action, most would expect intuitively that it would be and Ginet (though the arguments *against* him are not particularly persuasive) has not himself given a persuasive argument for believing otherwise. I would argue that what really matters in such scenarios – if we are to attribute ultimate responsibility (UR) and thus attain the Gold Standard freedom will – is where the *desire* came from in the first place, and Ginet up to this point seems to suggest that it arose essentially out of nowhere and with no explanation for *its* content. Is coming to have the desire itself an action? And if so, is it causally disconnected from its origins? There is no real discussion on these points and in any case the looming regress here is obvious.

Ginet's conditions do not therefore appear to be sufficient for the truth of the corresponding reasons-explanations and he offers no real positive account for why his version is preferable to an arguably more intuitive causal picture. But even if we were to accept it as Ginet presents it in the discussed examples, it does not address the key issues raised, such as how we can meet the important UR condition.

2.5.3 Jessica, Luck, and the Ever-Looming Regress

In order for an act to be free according to Ginet, then, it needs to be undetermined, initiated by a basic mental act – which is accompanied by an actish phenomenal quality – and be explainable in terms of the agent's reasons, by way of the presence of a concurrent intention which directly links the antecedent desire to the ultimate action.

Returning to our example of Jessica the would-be Olympian: as she sits, with her hand on the car door handle, Jessica has the desire to help the fallen man. Correspondingly (but in no way causally) the appropriate volition (a simple mental act not causally necessitated by antecedent states and events) arises, along with its intrinsic actish phenomenal quality that confers true ownership of the act to Jessica. The antecedent desire to help generates a concurrent intention (which Jessica retains as she acts) to open the car door and step out, which contributes directly to her satisfying her desire to help the injured man. Jessica's desire to help the man does not cause her to act but, combined with the concurrent intention held throughout the act of getting out of the car, is sufficient for the truth of an explanation for her action in terms of her reasons.

Ginet's non-causal approach, it could be said, asks us to accept some rather peculiar claims and (most importantly) does so without really providing in return a convincing positive account of libertarian action that advances the cause for such theories. Jessica's action (in the non-causal re-telling of her situation) is uncaused by her desire but, even if we were to grant Ginet's account, it is not clear how a causal break in that region of linked events would necessarily aid her free will when the formulation and linkage of the desire itself goes unexamined; seemingly leaving UR (and the hopes for the Gold Standard free will) in doubt. Ginet's main driving force in the development of his theory (in particular in its subsequent defence) appears to be the desire to refute the notion that uncaused actions cannot be ones that an agent performs for a reason; and this is because, on his account, it is only uncaused actions that can be considered free actions and it is only free actions that an agent can be held responsible for. If he fails to demonstrate that uncaused actions can be done for reasons, he is left in the difficult position of defending the notion that the only acts we can be held responsible for are ones which we do for *no* reason. But how does his theory fare against the key objections to libertarian theories? It is to this question which I shall now turn.

The Luck Objection

Does Ginet's non-causal approach have any advantages over Kane's event-causal approach in the face of the Luck Objection? A brief consideration of the objection shows that it does not appear to. As a reminder, the following restates the objection as it relates to the Jessica example. As she sits poised, with her hand on the car door handle, about to help or not help the fallen man, Jessica is not determined to do either by antecedent conditions so can follow either path – whatever her past circumstances. But if there is nothing within herself (within her CEP) that controls which outcome comes to pass, what she does ultimately choose to do can be only down to luck. In Kane's event-causal theory Jessica's psyche splits in two with one desire-driven, causally-produced (*by her CEP*) neural stream designed to cause her to drive on to her race and another equally desire-driven, causally-produced (*by the same CEP)* neural stream designed to cause her to get out and help. One neural stream ultimately wins out over the other by achieving some kind of activation threshold first and Jessica is caused to, say, get out and help. Her desire to help won the day and then caused her action to get out and is therefore the *reason* for which she acted as she did.

The first issue that arises when attempting to compare Ginet's approach is that his theory is broadly silent on the formation of the desires themselves and he generally does not go into any detail in his work when it comes to "torn" decisions of this type. The protagonists in his own examples (and those of his critics when discussing his work) tend to be working only with desires to perform a single action for different reasons, rather than competing and opposed desires which would lead to the performing of potentially vastly *different* actions. It seems fair to assume, in these cases, that the morally important matters of conscience occur in the forming of the *desire* not in the performing of the action and, without the agent being able to also exert some form of controlled influence over the forming of the relevant desires, the UR principle would be called into doubt and we would be left considering the limited concept of free *action* and not the deeper notion of free *will*, which we have labelled the Gold Standard. A further problem with Ginet's account is that – even if we could be satisfied as to the formation of the desires – with the imperative that the initiating action be uncaused, whatever desire is ultimately formed, the agent appears free to perform any number of alternative actions regardless of the current desire. Jessica, therefore, may be poised in her car and hold a desire to help the man, but which volition actually arises (to help or to drive on) is undetermined. If a volition to help is enacted (accompanied by the actish quality) and under the guiding light of the relevant concurrent intention, Jessica, on Ginet's view, acts freely, with control and can give a satisfactory reasons-explanation for her action. But what if, despite the desire and concurrent intention to help, a volition to drive on is enacted; a possibility which it seems must remain available due to Ginet's own brand of libertarian convictions. If the formed desire has no tangible, linked influencing factor over the action, then surely we are back to luck (or perhaps randomness) as the apparent arbiter of our deeds.

Another issue is, in regards to Davidson's objection to non-causal accounts of reasons-explanations, it could be argued that even if Jessica were to hold the desire to help the man because she believes it is the right thing to do, it could also be possible for her to be acting for a different reason that she also holds – say, that the man looks well dressed and she believes there could be a financial reward in it for her. Whatever her reasons, the fact that it seems difficult on Ginet's approach to account for the precise, tangible role reasons play in action can only play further into the hands of those who point to the Luck Objection as a decisive problem for libertarian theories.[31]

Rather than offering any advantages in dealing with the Luck Objection, therefore, the non-causal approach seems actually worse off than the event-causal approach. In Kane's event-causal theory, though there is arguably some inevitable arbitrariness in the formation of the desires (if they are not generated deterministically) and in the battle of competing desires, once one neural stream wins out over the other the action flows causally from the victor. In the non-causal approach, not only is there the same potential arbitrariness in the formation of the desires, but then there is a further level of arbitrariness at the volitional stage, which could seemingly produce a number of different courses of action – despite the desire that is ultimately held – as the mental act initiating the physical action is entirely disconnected.

The Basic Argument

Ginet believes his non-causal approach halts the regress which dogs all libertarian approaches, as it breaks the causal chains which inevitably generate this regress, and thus he would likely view his approach as superior to the other libertarian theories in the face of Strawson's Basic Argument.

> If voluntary exertion of the body involves the subject's causing her body's exertion by means of mental action, then voluntary exertion is not a kind of action that counters any of these theses (1) to

[31] Another line of argument is that many reasons never explicitly enter into our intentions in this way but still influence our actions. Given the confused and convoluted nature of the human psyche it is surely much more likely that a plethora of reasons could play a role in causally influencing our actions without explicitly entering into our intentions. See Clarke 2003 for more on this point.

> (3).[32] The regress does not stop with it. But it does stop with the mental action, the volition, that is the beginning of any voluntary exertion. Such mental action counters thesis (1), that acting always consists in causing something, because it is simple mental action, without internal causal structure, and any simple mental act is an exception to the claim that acting consists in causing something. (Ginet 1990, p.11)

Because the basic mental action – the initiating volition – is uncaused, Ginet is arguing, it is causally disconnected from the antecedent desires or motivations and therefore avoids an infinitely traceable regress. Jessica's act of getting out of her car, though accompanied by a remembered desire and concurrent intention, is not caused by these antecedent states so, the argument goes, is not inevitably generated from a CEP that she cannot be held fully responsible for forming.

There are two main issues to consider here. Firstly, if Ginet's theory is successful in breaking the regress (as he believes it is) this would seem only to increase his theory's susceptibility to the Luck Objection because (as discussed in previous sections) it is hard to understand how an action completely disconnected from an agent's CEP can be considered anything but a random occurrence (an example of the recurring *Catch-22 of Libertarianism*). Ginet may, of course, claim that such an objection only begs the question in favour of causal accounts and that the presence of a concurrent intention based on her desires is sufficient a link to prevent the act being considered as random. However, as O'Connor and Kane both point out,[33] referral and remembering is not the same as a direct (physical/tangible/ontological) link between the agent's reasons and consequent decision/action, and without such a link (causal or otherwise) there would seem to be nothing that could play the role of a kind of "appropriate action generator." As the causal approach is the most intuitive, the onus is on Ginet to provide a more ontologically satisfying link.

The second issue is that, if Ginet is truly claiming that a concurrent intention is a satisfactorily explanatory link and the concurrent intention is generated by the antecedent desire, then really he is

[32] The theses Ginet is referring to here are statements he introduces to demonstrate the different possibilities for where libertarians are going wrong, in regards to generating the regress. "At any rate, it is clear that at least one of the following three theses must be wrong: (1) the thesis that action consists in causing something, (2) the thesis that the *event*-causation analysis always applies to a *person's* causing something, or (3) the thesis that when a person causes something, the cause event must be an action. (Theses [3] entails thesis [2], but not the converse.) For these three together do yield the unacceptable regress." (Ginet 1990, p.7)

[33] See O'Connor 2000 and Kane 2005.

just shifting the regress from the point of action to the point of desire formation. The desire (by way of the intention) is the explanation of the action and so the *reason* for why it occurs. But where has the desire come from? If Jessica's desire is to help the fallen man, the libertarian case is not helped by this desire arising randomly from nowhere. But if it arises from Jessica's (historically developed) CEP then the return of the regress is inevitable and Ginet's non-causal view fares no better against the Basic Argument than the event-causal approach.[34]

2.5.4 The *Criteria*

Whilst an approach being seemingly less intuitive than its competitors is certainly not by itself a reason not to take such a view seriously, it does tend to put the onus on the proponents of such a view to proffer a clear justification for why such a theory should be accepted. Ginet, however, does not seem to offer much by way of a positive account for the structure of how his non-causal approach functions, contenting himself mostly with explaining that if his approach has inherent problems, then the other libertarian approaches do not fare much better. In any case, even if we were to accept Ginet's non-causal approach as a viable alternative based on its merits, the question is: does it come any closer to being able to be considered coherent? I shall now judge it against the *Criteria* in order to assess this.

The first three principles are broadly standard for all libertarian theories so can be taken together.

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.
P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.
P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.

Ginet states explicitly in his work that he believes free will exists and that it is incompatible with determinism and thus his theory clearly does rest on a platform of indeterminism and includes agents

[34] There is also the question over whether the formation of her desire (or reaching a decision to act) is itself an *action* in Ginet's use of the word. If so, further questions arise as to whether the desire to form a desire to help or not requires its own concurrent intention and causally disconnect action, which then results in a desire. See Ginet 2008, p.232 for part of Ginet's and Clarke's debate on this point.

that are free to act despite the antecedent states of the world.[35] It is therefore my contention that these first three principles have been met.

P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.

This is a difficult one to judge as there is generally little discussion concerning the decision-making process. However, as stated above, the options available to Ginet here do not look promising. Given that it is unclear how the guiding desires are formed, coupled with their disconnect from the subsequent action, it is difficult to see how Ginet can fully claim to have escaped either determinism *or* arbitrariness. If we are to offer the benefit of the doubt with regards to determinism, we are left with actions that can seemingly only be considered random as there is no explanation for why an agent *chooses* to follow one course of action over a viable alternative and, furthermore, this seemingly arbitrary choice is then followed by an action which is not connected to it in an easily understandable way. It is difficult, therefore, to see how Ginet's agent could claim full control over her acts and thus principle four has not been met.

P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.

In the context of non-causal theories, it could be argued that this principle is question begging from the start. In Ginet's theory the agent is not the cause of herself; the agent is the cause of nothing at all. However, the goal to escape the looming regress remains a key one for all libertarian approaches aiming for the Gold Standard free will I have identified and Ginet's theory fails to achieve it as the regress simply shifts from the disconnected action to the prior reasons/desires guiding the action. Principle five has therefore also not been met.

[35] See Ginet 1990, p.90, as one example.

P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

In regards to the final principle, it could be argued that removing the requirement for any causal connection removes any need to develop obscure forms of the causal relationship, of which both event-causal and (in particular) agent-causal theories have been accused of doing.[36] However, if the result of such a move is a theory which is just obscure generally, we do not seem to be in any better position. Despite all its flaws, it could be argued that the causal picture is more readily understandable. If non-causalists such as Ginet are to gain any kind of acceptance for their position, therefore, they perhaps need to develop a more detailed account of the ontological relationship between an agent's CEP and her actions which could be considered a satisfactory alternative to the causal picture. In light of the absence of any such account, principle six cannot be deemed to have been met.

2.5.5 Conclusion

In addition to being arguably metaphysically obscure and somewhat counter-intuitive, non-causal accounts such as Ginet's seemingly fail to solve any of the problems facing libertarian theories that I have outlined and cannot therefore be considered coherent based on the *Criteria*. As such, no good reason has yet been provided for deviating from the more standard, causal picture and it is therefore perhaps within that framework that the search should continue for an intelligible libertarian theory.

[36] More on this in the next section.

2.6 Agent-Causal Theories

Agent-causal theories seek to solve the problems that beset libertarian attempts to achieve their Gold Standard freedom of the will by taking an ostensibly more direct approach to causation by an autonomous agent; with the agent having full, free, and direct control over their choosing of a course of action from amongst genuinely available alternatives. In such a way, it is argued that the agent directly causes an action (or the coming to be of an intention to carry out an action). The theory's supporters claim that the approach aligns most with lay attitudes towards freedom and responsibility, being the most instinctively (and pre-theoretically) intuitive of the libertarian theories, as it presents a seemingly straightforward picture of an autonomous and enduring agent exercising absolute control over her future.

The notion of *control* is key for agent-causalists, as the main criticism they have of the competing libertarian theories is that they simply do not afford the agent any (or at least not enough) *genuine* control over their choices and actions. The event-causal approach, they argue, offers a weak form of control at best because, despite allowing for alternative possibilities (despite meeting the AP condition), it provides no satisfactory mechanism for how an agent can ultimately select from between them and thus directly bring about one course of action over another. An agent's reasons on the event-causal view can be key causal factors in cognitive processes which decide her action but these processes can only cause actions non-deterministically, meaning competing desires could – under the very same circumstances – cause different actions entirely. What is missing (the agent-causalists argue) is the agent taking a *direct* command over the selection process, without which there can be only probabilistic outcomes and an unacceptable level of randomness in the decision-making/action process. Non-causal theories, they argue, offer *no control whatsoever* as they have at their centre simple mental actions that are not caused by anything at all and, if the agent is not in any way determining such actions to occur, then they cannot be under her control. Proponents of agent-causal theories thus see the agent in both event-causal and non-causal theories as generally passive and incapable of truly making something happen, when what is actually needed is an autonomous agent that takes full control and actively *selects* and then directly *brings about* a course of action.

In order to try and improve the situation for libertarians, agent-causalists argue for a different kind of causal relationship between the agent and her actions (or intentions) than the familiar one of states or events (such as an agent coming to recognise certain reasons) causing further states or events (such as acting on said reasons). The causal relationship involved when an agent is acting freely is not reducible to the usual type of event-causation and such free actions are thus not caused by prior circumstances. They are instead caused only by *the agent*; taken to be an enduring substance, itself not reducible to events of any kind. This direct causation of an event by an enduring agent (and substances more generally) is often seen by agent-causalists as more fundamental than (or at the very least ontologically equivalent to) its event-causal cousin. It is a primitive causal relation: an ontologically basic causal capacity not further explicable in terms of causation by events that all autonomous agents must possess if they are to act freely and be considered ultimately responsible for their actions. Agents cause events immanently and not by doing anything else and thus are truly seen as *prime movers unmoved* – creators/initiators of new causal chains who cannot themselves be caused to create such a chain by anything else.

As Roderick Chisholm, one of the pioneers of modern agent-causation theories, puts it:

> If we are responsible… then we have a prerogative which some would attribute only to God: each of us, when we act, is a prime mover unmoved. In doing what we do, we cause certain events to happen, and nothing – or no one – causes us to cause those events to happen." (Chisholm in Watson (ed.) 2003, p.34)

It is because of her status as an uncaused cause of free actions – who can exercise a fully determining control over the relevant selection process – that an agent on the agent-causalist picture is believed by the theory's supporters to possess a required level of control that her counterparts in the other libertarian models simply cannot match; and which allows her to escape from the *Catch-22 of Libertarianism* as, despite her own determining capabilities, such an agent still remains – herself – undetermined to act in any particular way by antecedent conditions or prior circumstances. The agent directly causes events, but is not herself an "event" which could be caused.

Such considerations, however, lead to a number of key criticisms which are often levelled against agent-causal theories; one of the most powerful being that the resultant models for free action – which feature enduring agents exercising a type of irreducible substance causation – rest on seemingly contentious ontological commitments. Also, whilst a direct form of causation from agent to action might seem ostensibly to provide for greater agential control, it is not immediately clear how – if the agent herself is not caused to act by anything at all (including relevant antecedent states such as beliefs or desires) – we can account for exactly *why* an agent chose to “agent-cause” a particular act in the first place. As with non-causal accounts, there is a difficulty in providing a satisfactory reasons-explanation for an agent’s acts because, if the decision cannot be caused by the agent’s antecedent states, but only by the enduring agent themselves, then how are we to make sense of it in terms of her reasons for doing it? Does the addition of such a *prime mover*, therefore, actually do anything to rid us of the arbitrariness apparently inherent in libertarian theories? Indeed, are there in fact any advantages to the agent-causal approach that would make accepting its likely contentious ontological commitments worth the price? Such questions are what I propose to consider in the next section.

2.7 Timothy O'Connor's Causal Powers Approach

Despite an apparent lack of widespread support for agent-causal theories in recent times, a number of sophisticated defences of the approach have still been developed and one of the most advanced is that proposed by Timothy O'Connor. It is his "causal powers" approach that I shall focus on in this section (on the assumption that the broad criticisms discussed would apply to most other variants).[37]

O'Connor makes it clear in his work that he subscribes to the general agent-causal view that both event-causal and non-causal theories do not provide for a sufficient level of agent control in action and decision-making; as he explains in *Persons and Causes* (2000):

> A prima facie problem for this [event-causal] position is to explain how the agent directly controls the outcome in a given case. There are objective probabilities corresponding to each of the possibilities, but within those fixed parameters, which choice occurs on a given occasion seems, as far as the agent's direct control goes, a matter of chance. (O'Connor 2000, p.xiii) [38]

O'Connor therefore aims to demonstrate that agent-causal theories can overcome this problem of insufficient control, as well as flesh out his own particular model in such a way as to address accusations of metaphysical exceptionalism that critics have levelled against the position. As such, his task is largely a defensive one and he breaks it down into three main parts: an argument for the acceptance of an irreducible (non-Humean) causation; an account of how agent-causal actions can be explained in terms of the agent's reasons (which can in no way *cause* such actions); and a discussion of how the agent-causal capacity might be realised (physiologically speaking) in human beings.

[37] One possible (partial) exception could be Randolph Clarke's own "integrated agent-causal account" which attempts to combine indeterministic event causation (with accompanying reasons-explanations) with the agent-causal capacity. Though this may be seen as an improvement in the sense of bringing agent-causal views back towards more conventional thinking, it still fails against the *Criteria* for broadly the same reasons – with the addition of suffering from the event-causal issues *as well as* the agent-causal ones. See Clarke 2003 for more.

[38] Whilst this particular quote focuses on event-causal theories, O'Connor also makes similar points in regards to non-causal approaches in his work. Despite reasons not being able to cause an "agent-causing" there is an important "control creating" causal-relation in O'Connor's theory which he argues is missing in the non-causal theories. More on this in what follows.

2.7.1 A New Species in the Causal Genus?

Similarly to agent-causal theories that have preceded his, O'Connor sees (Gold Standard) free will as based on an ontologically irreducible causal relation between an agent (as an enduring substance) and events internal to her action, and is thus well aware that if an argument for such an irreducible causation *cannot* be made then agent-causal theories are dead in the water. His first task, therefore, is to make such a notion as palatable as possible.

> I begin with a strong, highly controversial assumption about the general concept of causality. This assumption is that the core element of the concept is a *primitive* notion of the 'production' or 'bringing about' of an effect. This entails the negative thesis that a satisfactory reductive analysis of causality along Humean lines (in any of its versions) cannot be given. It should be readily apparent that if, contrary to this anti-Humean assumption, a satisfactory reductive analysis of causality *can* be given the agency theorist's project of defending a variant species of causality immediately collapses into incoherence. For such reductive analyses are either committed to a general connection between certain *types* of causes and effects or equate causation with a form of counterfactual dependence. Neither approach is consistent with the agency theorist's claim that a causal relation can obtain between an agent and some event internal to himself, since his understanding of this is such as not to imply that the sort of event effected on that occasion will or would always (or generally) be produced given relevantly similar internal and external circumstances. (O'Connor in O'Connor (ed.) 1995, p.175)

As opposed to the more familiar picture of a reductive event-causation, O'Connor is therefore looking to base his theory on this primitive idea of causal production: "(in a more technical jargon, "causal oomph")" (O'Connor 2000, p.67). Whilst not denying that event-causation exists, he argues that it just does not submit to a reductive analysis and thus he instead chooses to base his model for event-causation on what he calls the older and more traditional "causal powers" account developed by R. Harre and E. H. Madden (1975), in which certain objects produce characteristic effects in particular circumstances due to the properties they possess and the causal potentialities they can therefore manifest.

> When placed in the appropriate circumstances, an object manifests its causal powers in observable effects. An object's powers are based in its underlying nature, for example, its physical, chemical, or

> genetic constitution and dynamical structure. Circumstances prompt the exercise of a power in one of two ways: either by stimulating a latent mechanism to action or by removing inhibitors to the activity of a mechanism in a state of readiness to act. (O'Connor 2000, p.71)

Having argued that such a non-reductive *event-causation* should be a perfectly acceptable causal model, his next step is to argue an equivalent model for *agent-causation* can be understood in the same way and the latter should therefore not be viewed as any more mysterious than the former. Both variants share the same primitive feature of "causal oomph," as he puts it, but differ in that – for agent-causation – the first relatum is not an antecedent event but the agent herself. With this addition of an agent-causal variant of such causal power, which O'Connor (following Reid) calls "active power," O'Connor is attempting to move away from a physical framework in which it is only events that can cause other events, as (to his mind) it is this type of framework that does not allow for a truly autonomous and controlling agent as such event-causings must be either wholly determined by other, antecedent events or be subject to solely probabilistic outcomes. It is the ability to exercise this "active power" that confers the true, Gold Standard free will on an agent.

O'Connor acknowledges that agents are somewhat unique in the way that they are able to exercise such an "active power" in causing events internal to themselves – such as coming to have an intention to carry out a specific act, which can also be seen as coming to a decision in the face of competing and contradictory possibilities. Such a causing is seen as the agent's basic action, which then initiates an event-causal chain to carry out the specified action in a similar fashion to Ginet's non-causal model.[39] The important difference (for the agent-causalist) is that the agent directly *causes* this basic mental action, which O'Connor sees as key to ascribing ownership and responsibility. Agent-causation then (unlike event-causation, even on the same causal powers model), cannot be understood as a "function from circumstances to effects", but relies on objects with certain properties manifesting causal powers under appropriate circumstances.

> [P]arallel to event causes, the distinctive capacities of agent causes ('active powers') are grounded in a property or set of properties. So any agent having the relevant internal properties will have it directly within his power to cause any of a range of states of intention delimited by internal and

[39] And, as with Ginet, this notion of a "basic action" or "simple mental act" is an important part of the "free" decision-making process as it is seen as an action which an agent can perform, without them needing to perform any additional action as a means to do so.

> external circumstances. However, these properties function differently in the associated causal process. Instead of being associated with 'functions from circumstances to effects,' they (in conjunction with appropriate circumstances) make possible the agent's producing an effect. These choice-enabling properties ground a different type of causal power or capacity – one that in suitable circumstances is freely exercised by the agent himself. (O'Connor 2000, p.72)

The agent's activity is therefore in no way simply a product of external conditions acting on her internal states; she is the sole originator of her own actions. Via this mechanism, and due to the presence of properties which enable but in no way necessitate (or even make probable) a choice, the agent can exercise her active power at will in a direction of her choosing in order to cause a particular kind of event within herself – the coming to be of a state of intention to carry out a particular act, which functions as a choice between competing alternatives.

Whether or not O'Connor's version of such a substance-causal picture (or indeed any version of it) can be successfully argued for as ontologically more basic (and thus the more fundamental) than the reductive event-causal alternative is a widely debated prospect, and one which I do not propose to delve further into here.[40] What is most important for my discussion on the general intelligibility of libertarian theories is whether accepting such a "causal powers" account of causation – and incorporating it into a libertarian model of free will – would in any way improve the libertarian position in regards meeting my coherence criteria and escaping the *Catch-22 of Libertarianism*. Does the addition of a such a wholly determining, but entirely uncaused, agent-causal power actually result in an agent who can make decisions completely non-arbitrarily and thus directly control their actions, without themselves being in any way subject to antecedently determining conditions?

[40] Several arguments have been raised against O'Connor's account and its previous incarnations. One key one is that the approach is contradictory due to that fact its defence of agent-causation whilst advocating a causal powers account of event-causation leads to a singularist verses antisingularist contradiction. Singularism, which is implied by O'Connor's account of agent causation, states that causation is a singular rather than a general matter where the causal connection between two events does not depend on anything extraneous to that relation and has no implications for any other event happening at another place or time. The causal powers account of event causation, however, is seen as an antisingularist account, which views causal relations as part of more general patterns and thus posits general causal laws. O'Connor does himself acknowledge he could be accused of trying to "have things both ways" (O'Connor 2000, p.72). For more on this see Hiddleston 2005, p552 and O'Connor's defence (O'Connor 2000, p.72).

Jessica: A Prime Mover Unmoved?

If our athlete Jessica were to possess such an agent-causal capacity, we might envision her dilemma as follows. As she sits, with her hand on the door handle, surveying the fallen man and despairing over the prospect of missing out on potentially life-changing opportunities, she weighs up her options and the pros and cons of pursuing each one. All the relevant information presumably gets filtered through her current established psychology, CEP, as she deliberates and then at some stage the agent-causal capacity takes over and *decides* which course of action she will follow, simultaneously directly causing the coming to be of an executive intention to, say, drive on. This basic action then triggers an event-causal chain which culminates in her doing just that. Thus, she drives on leaving the man to his fate. This agent-causing of her intention to drive on is not itself determined by any antecedent factors, nor is it the subject of an indeterministic causation by antecedent events; but neither is it uncaused entirely. It is caused by *her*, directly, as the controlling agent. Jessica, the agent-as-enduring-substance, steps in and causes her decision to drive on, which she is capable of doing due to her ability to exercise her “active power,” generated from having the relevant properties in conjunction with appropriate circumstances. Any antecedent or concurrent (presumably event-causal) processes taking place within her CEP do not in any way cause this agent-causing, either deterministically or probabilistically. Yet, what *is* caused (the decision to drive on) is *determined* to occur, but determined by the agent and not any antecedent factors or anything else.

The obvious difficulty with this picture (in addition to any possible ontological unconventionalities) is that – as with the non-causal approach – we seem to have a clear disconnect between the basic action of causing the coming to be of the intention and any antecedent or present motivations (such as reasons, desires, or beliefs), leaving us requiring an account for how such motivations can influence our decisions. Unlike on the event-causal view, reasons cannot be events that (even non-deterministically) play a role in causing an agent’s free decision, because such a decision cannot be caused by anything except directly by the agent herself. Yet, in order for us to properly ascribe responsibility, it would seem an agent must be acting on the basis of her reasons. On O’Connor’s account, how can we explain the relationship between the agent’s reasons and her ultimate action? If it is not a causal relation, then what is it? O’Connor makes efforts to address the issue, which I shall consider next.

2.7.2 Non-Causal Reasons

As with proponents of the non-causal approach, then, agent-causalists seem to be confronted with a problem when it comes to giving an account of how the actions of an agent can be explained in terms of the reasons she may hold for performing it. As just discussed, on the agent-causal view a decision cannot be caused by anything except directly by the agent herself – a primitive, ontologically basic enduring substance not reducible to any *event* or *series of events*. Reasons, therefore – taken as mental events – cannot seemingly be a cause of the agent's decision or action. The questions which then arise are the same as those we considered in the previous section when considering Ginet's non-causal account. If the relationship is not a causal one, what is the nature of it? What influence can reasons, beliefs, desires and so forth have on agent-causings, and thus the decisions and actions of the agent?

As already mentioned above, an agent holding reasons for acting as she does is normally central to all theories of freedom and responsibility, both compatibilist and indeterminist. In the compatibilist arena this is a relatively simple endeavour, with reasons forming part of the antecedent, determining conditions which ultimately result in a specific decision and/or action. For indeterminists, though, we have the ever-problematic *Catch-22 of Libertarianism* with, on the one hand, a requirement for decisions and actions to be products of the agent's CEP (which would include her reasons) in order to avoid charges of arbitrariness and, on the other hand, the important proviso that such reasons must not (necessarily) be antecedently determining – seemingly circling us back to charges of arbitrariness. In event-causal accounts, such as Robert Kane's, reasons *are* causally influential (at least in an indeterministic fashion) but even this is not allowed on O'Connor's account because, as already stated, *nothing* can cause an agent-causal event, indeterministically or otherwise. Also, in O'Connor's opinion, such indeterministic causation would in any case not confer sufficient control upon the agent for a satisfactory ascription of responsibility.

O'Connor's own particular account of how reasons *are* able to influence free choices without being a direct cause of agent-causal activity has a number of parts to it, which he has developed and added to in response to various criticisms. At its core is a schematic similar to that employed by Ginet, which attempts to explain actions solely in terms of the agent's prior desires. Both positions deny that reasons can explain an action only to the extent that they contribute to directly *producing* it, but O'Connor adds in the key component that the agent nonetheless directly causes the relevant intention.

> The agent acted then in order to satisfy his antecedent desire that Θ if
>
> 1. prior to this action, the agent had a desire that Θ and believed that by so acting he would satisfy (or contribute to satisfying) that desire;
> 2. the agent's action was initiated (in part) by his own self-determining causal activity, the event component of which is the-coming-to-be-of-an-action-triggering-intention-to-so-act-here-and-now-to-satisfy-Θ;
> 3. concurrent with this action, he continued to desire that Θ and intended of this action that it satisfy (or contribute to satisfying) that desire; and
> 4. the concurrent intention was a direct causal consequence (intuitively, a continuation) of the action-triggering intention brought about by the agent, and it causally sustained the completion of the action. (O'Connor 2000, p.86)

Thus, according to this explanation, having a desire to do *A*, and holding it concurrently with (causally) initiating the basic action to fulfil the desire to do *A*, and throughout the actual doing of *A*, are sufficient conditions for a satisfactory association of the desire to *A* with the doing of *A* – whilst not admitting of any direct causal relationship between the desire and the agent's "self-determining causal activity." However, O'Connor does then go on to bolster this picture by affording the prior motivations an ability to have an attenuated form of causal influence by way of the recognition of relevant reasons within the agent increasing an "objective propensity" for her to cause an intention to act. Such activity, then, can be said to causally *structure* the agent-causal capacity by affecting the probabilities of the agent causing certain intentions. And "prior motivations" here does not just include proximal reasons but also longer-term states of character and dispositional attitudes and so forth.

> [P]erhaps we can achieve the desired causal connection between reason and action by supposing instead that the agent's coming to appreciate a reason for acting appropriately affects (in the typical case, by increasing) an *objective propensity of the agent* to cause the intention to so act. On this latter suggestion, while nothing produces an instance of agent causation, the possible occurrence of such an event has a continuously evolving, objective likelihood. Expressed differently, agent causal power is a *structured* propensity towards a class of effects (the formings of executive intentions), such that at any given time, for each causally possible, specific agent-causal event-type, there is a definite objective probability of its occurrence within the range (0, 1), and this probability varies continuously as the agent is impacted by internal and external influences. Where the event

> promoted occurs, the effect of the influencing events is to alter the prior likelihood of an outcome, not to produce it. (O'Connor 2009, p.120)

Further to this addition of structuring causes (and in answer to Davidson's argument that only causal relations can distinguish between potential and actual reasons for an agent's acts; as discussed previously in section 2.5.2), O'Connor emphasises that there is a distinct type of content to the agent-caused intention which is intended to ensure that it is specified from the outset precisely *which* reason is being satisfied. Whereas I may sometimes act generally based on reasons I hold (consciously or unconsciously), there are other occasions where I might expressly act in order to achieve a particular goal that would satisfy a particular reason (or set of reasons) I have for achieving that goal. In the former situation I could be said to be acting *on* a reason (which is a structuring cause) but in the latter situation I would acting *for* a reason (a stronger notion) and in such a case I am conscious of the reason and the goal enters into the content of the intention. What this does, O'Connor argues, is take things a step further than just relying on structuring causes and guarantees which reason is the *actual* one (or ones) for which I am acting. The content of the intention is that the agent performs action X *in order to satisfy reason Y*.

> Generally, I am conscious of certain reasons that favor the course of action I am choosing. And sometimes, I expressly choose the action for the purpose of achieving the goal to which those reasons point. That is, this goal enters into the content of the conscious intention I form. In such cases, instead of simply intending to *A*, for some action type *A*, I cause the intention to *A for the sake of G*, where *G* is the goal of a prior desire or intention that, together with the belief that *A*-ing is likely to promote *G*, constitutes the consciously grasped reason *for* which I act. Now, since I freely and consciously bring the intention into being and thus give it just this purposive content, that purpose cannot but be one for which I am acting. What is more, a further explanatory connection between that reason and the choice is forged beyond the reason's influence on the choice's prior probability. This connection consists in the conjunction of the external relation of prior causal influence and the purely internal relation of sameness of content (the goal *G*). There may be several reasons that increase the likelihood that I would cause the intention to *A*. In the event that I do so, each of these reasons are ones *on* which I act. But if I am conscious of a particular reason, *R*, that promotes a goal *G* (and no other reason promotes that goal), and I cause the intention to *A* for the sake of *G*, then *R* plays a distinctive explanatory role, as shown by the fact that it alone can explain the goal-directed aspect of the intention's content. It alone, as I shall say, is one *for* which I act. (O'Connor 2009, p.121)

Expanding further on Jessica’s cognitive process, then, we can add in a more specific role for her relevant motivations. She will likely have reason-based competing desires which will inform her choice as to what course of action to pursue. Given her lifelong commitment to competitive athletics and a presumably fierce desire to succeed in the sport, we can envisage such strong motivating factors would qualify as O’Connor’s “structuring causes” which may then increase the *objective propensity* that she will cause the intention to drive on to her race and leave the elderly man to his fate. Furthermore, should she choose to do just that, the more proximal reasons (such as “I desire to win this race in order to fulfil my lifelong dreams”) of which she is well aware will feature in the content of the intention. Jessica will therefore be causing the intention to *A* (*to drive on*) for the sake of *G* (*being able to compete in her race*), with reason *R* (*I desire to win this race in order to fulfil my lifelong dreams*) providing a satisfactory explanation for the decision. But is such an explanatory link sufficient to counter Davidson’s objection that reasons must themselves play a direct causal role if we are to be able to account for what it is that makes a particular reason *the* reason for which an agent carries out a particular action?

O’Connor’s initial schematic (quoted above), it will be noted, is very similar to Ginet’s non-causal account of reasons-explanation discussed in section 2.5.2 above, relying, as it does, on the presence of a concurrent intention which also has prior motivations as part of its content. As such, it is vulnerable to the same objections and Ginet (and other non-causalists) therefore argue that, as the addition of an agent-causal capacity into their similar framework does not provide an answer to any of the criticisms levelled at their account, it is thus not only potentially metaphysically dubious but simply unnecessary.[41] O’Connor, of course, argues that the agent-causal capacity *is* adding something important by highlighting the fact that choices have a causal structure and are not simple mental events in the same way as presented by Ginet and others. The relevant intentional state does not just have as part of its content a specification of which reason is being acted upon, but is also *agent-caused* in a manner explicable by the reasons held. He also believes he is (*contra* Ginet) giving an explanation for *why* free actions are uncaused by antecedent events: because the first relatum is a substance, not an event, and thus by definition not something that itself could be caused by further events. It is of course open to Ginet to counter that this interpretation does not add anything extra to his view that the free actions simply are not caused *at all*.

[41] See Ginet 1990.

In any case, it is not clear how just the addition of an agent-causal capacity makes O'Connor's account any less susceptible than Ginet's to charges that without a direct causal influence we cannot achieve a truly satisfactory account of free choice in terms of the agent's reasons for acting. As discussed when considering Ginet's approach to reasons-explanations, it is objected that merely having a reason for an action (even if it appears said reason is consciously recognised as the reason for which the agent seems to be acting) does not necessarily make that indisputably the reason *for* which an agent acts and simply stipulating that the reason forms part of the content of the intention and therefore must, by definition, be *the* reason for which a particular act is carried is not a convincing argument in favour of non-causalist over causalist reasons-interpretations. To begin with it seems far too simplistic an approach to take towards human cognitive activity, in which there is likely in reality to be all manner of competing (as well as differing but broadly aligned) motivations to carry out any act. Also, as Randolph Clarke has objected, having an intention refer to an action in this way means that what ultimately explains the action depends on the outcome of something posterior to it, which, he argues, seems absurd.[42]

O'Connor seems to acknowledge the potential weaknesses in this area – as evidenced by his attempt to strengthen his causal picture by affording reasons these structurally-causal abilities – and this addition, it could be argued, does indeed go some way towards giving reasons and other motivations a more satisfactorily explicable role in influencing free actions. But it does so seemingly at the price of adding in a significant event-causal component to O'Connor's theory. If reasons now probabilistically raise an objective propensity for the agent to act in a certain way, are we in a much different position to that of Kane's agent whose actions are indeterministically caused by her reasons, which O'Connor himself has criticised as not affording the agent enough control for a satisfactory ascription of responsibility?

On the one hand, O'Connor wishes to argue that agent-causings (by definition) are not and cannot be caused by anything else other than directly by the agent, which leaves him vulnerable to the same criticisms as the non-causalists when it comes to providing reasons-explanations for free actions. On the other hand, he wishes to argue that the agent-causal capacity can be causally-structured in a probabilistic sense by prior motivations – either proximal or enduring – which leaves him vulnerable to the same criticisms as the event-causalists when it comes to understanding how one possibility becomes actual over another. In either scenario, it is unclear why – given that little seems to be gained by adding

[42] See Clarke 2003, Chapter 8.

it – it is worth accepting that the addition of the agent-causal capacity is necessary. With the apparent ontological separation of the agent-causal capacity from the event-causal processes surrounding it, we retain the “explanatory gap” of *why* (in terms of prior motivations) an agent ultimately agent-causes a particular intention – as we have in both non-causal and (indeterministic) event-causal theories, though for different reasons. Only on O’Connor’s view, we have the *added* difficulty of the ontological status of the agent-causal capacity itself, arguments for the possible existence of which form O’Connor’s third main strand of his defence of the agent-causation approach.[43]

2.7.3 Is O’Connor’s Agent Any Less Lucky?

The key question for this section is whether the addition of an agent-causal capacity, such as that described by O’Connor, affords the agent any greater level of control in such a manner as would counter the Luck Objection. Does the agent’s ability to deterministically cause her actions by way of her (uncaused) agent-causings eliminate the problem of randomness or arbitrariness that generally dog libertarian theories? It is not initially clear, on both counts, that it would. Even were we to accept that an agent (as an irreducible and enduring substance) is capable of initiating a wholly new event-causal chain by bringing about an intention with the correct content, as just discussed, there still remains a fundamental disconnect between any antecedent motivating factors and the agent’s agent-causing a decision. This disconnect featuring immediately prior to the agent enacting her (fully determining) agent-causal capability seems just to move the position of the inevitable arbitrariness one step back in the process; from being part of the formation of an undetermined decision, to being part of the formation of an undetermined agent-causing. Jessica and her alternate-universe-counterpart Jessica*, who are exactly the same up to the moment of decision (as are the worlds around them), may well both have the capability to agent-cause a particular outcome to her dilemma in a determined fashion, but there is apparently nothing to prevent them agent-causing completely different courses of action, despite the identical nature of their antecedent circumstances. Both Jessicas may be fully (and solely) determining their particular course of action, but which action they agent-cause remains undetermined

[43] Whilst the defence of the potential existence of such an agent-causal capacity is an interesting debate, I do not propose to delve into it here. As it is my intention to show that the addition of such a capacity solves none of the problems inherent in libertarian theories of free will, arguments for how such a capacity might be realised (physiologically speaking) in human beings is not necessary for the present discussion. Those who wish to learn more about it can read chapter 6 in O’Connor 2000.

prior to it being agent-caused and thus the theory is vulnerable to the Luck Objection in the same way as other libertarian theories.

O'Connor's answer to this is essentially to argue that anyone who thinks that just because the agent-causal capacity is itself uncaused then it must fall foul of the same objections that beset other indeterminist accounts has simply misunderstood the concept of agent-causation. The agent-causal capacity, as he sees it, is a primitive form of control and, as such, just as it *cannot* (by definition) be caused by prior events it *cannot* (by definition) occur randomly or by chance. Jessica's exercising her free will is *constituted by her agent-causing her decision*; employing the agent-causal capacity *is* the exercise of control and therefore there does not need to be any further (prior) causal activity to explain it.

> The agent causationist contends that both these objections fail to take seriously the concept of agent causation. It is conceived as a primitive form of control over just such undetermined, single-case outcomes. The agent's control is exercised not through the efficacy of prior states of the agent (as on causal theories of action), but in the action itself. Alice's causing her intention to tell the truth is itself an exercise of control. And because, ex hypothesi, it is literally the agent herself generating the outcome, it is hard to see how the posited form of control could possibly be improved upon. So wherein lies the luck? (O'Connor in Kane (ed.) 2011, p.324)

As mentioned earlier in this section, by simply defining his agent-causal capacity as something that can neither be caused *nor* be random, O'Connor (and other agent-causalists) seems to be arguing by stipulation. Or rather – as Kane points out when discussing the agent-causalist answer to charges their view has issues with determinism and/or randomness – by "double" stipulation.

> In response to the objection that for all we know immanent agent-causation might be determined by hidden causes, they insist that immanent agent-causation is not the sort of thing that could in principle be caused or determined by prior events or circumstances. Now, in response to the randomness and luck objections, they add that the agent-causal relation is not the sort of thing that could in principle occur randomly or by chance either, since it is the agent's consciously controlling something. (Kane 2005, p.51)

Alfred Mele, whose extensive writings on the problem luck and chance poses for indeterminist theories of free will we have discussed in earlier sections, actually grants that introducing the agent-causal capacity may well provide the agent with a greater degree of control, but that this does not aid the approach in its attempts to combat the Luck Objection as the problem of luck (as Mele poses it) is an inescapable problem for *all* indeterministic accounts of free will regardless. The key fact is that without there being anything about Jessica's or Jessica's* characters, beliefs, desires, states of mind, life histories and so forth (their CEPs) which explains why one agent-caused the intention to drive on and the other agent-caused the intention to get out and help the fallen man, it can always be argued that there is an element of chance or luck in the process. To put it another way, without there being available a contrastive explanation for exactly *why* an agent acted in one way rather than another at a particular time – given that the alternate course of action was also available based on the initial conditions – then we can only put the difference down to luck. And this is true regardless of whether or not it is allowed that whatever each of the Jessicas ultimately agent-causes is done so in a direct and determined fashion which confers greater agential control than either the event-causal or non-causal approaches can offer.

O'Connor's reply to this (in a similar vein to Kane) is to deny that the lack of availability of a contrastive explanation is important and that the availability of a *non*contrastive explanation, which references the causes, reasons or other motivations is sufficient to achieve the Gold Standard free will and any associated ascription of responsibility (so long as it is referring to the acts of an agent with the agent-causal capability). Indeed (again like Kane), O'Connor believes that the type of contrastive explanations Mele is claiming are required to avoid charges of luck are simply not going to be available for any indeterminist theory of free will, but that being able to explain why an agent did one act instead of another by referencing prior preferences or elevated objective propensities is satisfactory. There may be some form of luck or chance or arbitrariness involved, but it is not of a responsibility-undermining form and is made even less so by the addition of the agent-causal capacity.

> …it might be best after all for the agent causationist to dig in his heels over what is meant by 'luck' in 'the problem of luck.' It is his (my) position that, on event-causal libertarian theories (including DSL),[44] an agent controls which of a range of possible indeterministic choices is actually made in at best a highly attenuated sense. In a corresponding sense, which choice is made, on such theories, is a matter of 'luck' (good or bad as may be). We further contend that, within the context of a suitably

[44] Daring Soft Libertarianism is Mele's own preferred theory of free will (see Mele 2006).

> developed theory of the guiding role of reasons, the concept of agent causation captures the missing element of direct control, so that agent-causal theories are not plagued by any problem of luck, though it is consistent with a stipulative notion of 'luck' that is tied to the absence of conditions grounding certain kinds of contrastive explanations. (O'Connor 2007, p.160)

The problem for O'Connor here, is that he is insisting that on his theory the agent has a much firmer, direct, and unassailable control over her actions than other libertarian positions because, like on the compatibilist event-causal view, such actions are *determined* by the agent in the strong sense. Yet, the compatibilist account *can* offer a contrastive explanation, which arguably means O'Connor's theory should also be able to provide such an explanation. Its failure to do so could therefore be seen as him not quite meeting the goals he has set for himself.

As above, if all parties are agreeing that some kind of luck or arbitrariness is inescapable, then O'Connor needs to be convincing that his addition of the agent-causal capacity *is* indeed adding some kind of extra, luck-reducing control not offered by its libertarian event-causal competitors. It is not clear O'Connor has achieved this, however, calling into question the requirement for the agent-causal capacity he has described (as well raising questions of whether it likely exists at all). Furthermore, it can also be argued that by adding in the notion of indetermistically varying objective propensities which causally structure the agent-causal capacity and probabilistically influence the eventual outcome (by an event-causal process), O'Connor is undermining the very control he is seeking to gain by introducing the agent-causal capacity in the first place. Which indeterministic propensities are ultimately realised, it could be said, would seem largely a matter of luck, with the agent-causal capacity relegated to the position of slave to the probabilistic whim of the event-causal processes. O'Connor, of course, denies that the agent-causal capacity is in any way diminished by this structuring arrangement, arguing that whatever the final probabilities for action become, the agent (as a causally efficacious enduring substance) remains the true arbiter of her decisions. It is not the involvement of probabilistic causes *per se* that raises the spectre of luck for event-causal theories, but the lack of ultimate control over the final decision.

> In reply, the agent causationist will insist upon the importance of the distinction between (the persisting state or event of one's having) reasons structuring one's agent-causal power in the sense of conferring objective tendencies towards particular actions and reasons activating that power by producing one's causing a specific intention. Nothing other than the agent himself activates the

> agent causal power in this way. To say that I have an objective probability of 0.8 to cause the intention to join my students at the local pub ensures nothing about what I will in fact do. I can resist this rather strong inclination just as well as act upon it. The probability simply measures relative likelihood and serves to predict a distribution of outcomes were I to be similarly inclined in similar circumstances many times over (which, of course, I never am in actual practice). From the agent causationist perspective, the reason that the alternative, causal indeterminist view is subject to the luck objection is not that it posits objective probabilities to possible outcomes but that it fails to posit the kind of control needed directly to determine what happens in each case…The agent causationist's solution is to posit a basic capacity of just that sort, although allowing that the capacity is not situated within an indifferent agent, but one with evolving preferences and beliefs. (O'Connor in Kane (ed.) 2011, p.326)

Ultimately, it appears it can be argued that – in an effort to bring the agent-causal theory into closer alignment with its seemingly more intuitive and straightforward competitors – O'Connor ends up attempting to straddle both the event-causal and non-causal approaches; taking bits from both sets of theories. Despite this amalgamation, however, O'Connor's model does not manage to solve this recurrent problem of luck, chance, or arbitrariness. Rather, by incorporating both a (partly) non-causal approach to reasons-explanations and a (partly) causal-indeterminist approach to structuring factors, he leaves himself open to objections focussing on both. Arguing that the agent is, in the end, just a bit *less lucky* than those proposed by the event-causalists due to the agent-determining final stage in the process is unlikely to be convincing when the agent-causal capacity itself is difficult to accept and it is not clear, even if we were to accept it, that the way in which O'Connor has presented it does actually reduce the level of luck apparently inherent in libertarian theories.

2.7.4 Any Less Regress?

As with both Kane and Ginet, O'Connor believes his particular model for indeterministic decision making is enough to halt the type of regress Galen Strawson (and others) have argued is inevitable and can therefore counter the Basic Argument. Also like both Kane and Ginet, O'Connor's claim to success in this area is not entirely convincing. It is easy to see why O'Connor would feel his approach has the upper hand over the others when it comes to halting the regress as there is not only a (at least ostensibly) clearly visible causal break between the agent's prior motivational states and her action, but also an accompanying *determination* of the action (or at least the intention to so act) by the agent. This is

O'Connor's attempt to address (and get around) the *Catch-22 of Libertarianism*, which (as I have said before) essentially states that the only way for agents to be truly in control of their acts is by determining them – as otherwise chance must play too big a role in the decision-making process for us to satisfactorily ascribe responsibility – but, if their acts are determined by their CEP, we enter into a vicious regress which again leaves us in a situation where we cannot satisfactorily ascribe responsibility. O'Connor's answer is to bestow upon autonomous agents a fully determining causal capacity, which itself cannot be caused – thereby giving the agent the control they need without them slipping into the inexorable regress. The trouble is, as we have seen, O'Connor's structurally causal additions (which he is forced into including in order to ground his agent's actions within her CEP), combined with the apparent disconnect between such structuring causes and the ultimate agent-causing, seem only to admit a good chunk of arbitrariness into the process.

Does O'Connor's model fare any better against the other side of the Catch-22: the apparent regress made inevitable by the continuous picture of causal flow? O'Connor argues that it does, and its success comes from the distinct causal break in the proceedings (which is also, of course, where both Kane and Ginet argue the success of their theories in this area comes from) generated by the entirely uncaused, agent-causal capacity.

> let us consider what the agent causationist might say in reply to Strawson. Aware of certain reasons pro (*r1*) and con, I cause an action-initiating intention to A. This is explained by my having been aware of reason *r1* while deliberating and as I completed the action, a reason that increased the prior probability of my choosing to A. I did not directly choose to be in a state of being aware of and motivated by *r1*. I simply found myself in that state, among others, and proceeded to deliberate. The totality of such conative and cognitive states circumscribed the range of possibilities for me, and also presumably the scope of responsibility directly connected to my free choice. But that choice was neither fully causally determined by those states nor merely a "chancy" outcome of tendencies of those states. Instead, I directly determined which choice within the available range would be made. This choice is explained by "how I was, mentally speaking" at that time, but it is not fully the result of that state. (O'Connor in Kane (ed.) 2011, p.320)

In the attempts to remove the "chancy" elements of his model, O'Connor emphasises the causal role of reasons in raising the probability of a particular course of action being chosen and then (freely) enacted by the agent-causal capacity. As discussed above, I do not believe O'Connor is successful in removing the arbitrariness from the system but, were he to be so, the strategy he employs involves varying probabilities to act created by prior motivations, which is surely sending us towards a regress and putting his role in conflict with the condition of ultimate responsibility (UR). O'Connor is quite clear in the above passage that relevant reasons and the states of being aware of them arise entirely unbidden and, as discussed in section 2.2 on Kane's event-causal theory, it is surely open to us to conclude that, as such states play a crucial role in both prescribing the available possibilities for action and altering the probabilities for each possibility – plus presumably arise in a deterministic fashion (otherwise they would not be doing anything to *reduce* the level or arbitrariness in the system) – an important regress is still being generated.

O'Connor does acknowledge the issues here, and also notes the likely regressive impact of a whole host of "powerful and deep behavioural and attitudinal dispositions" which can often be tracked back through the life history of an agent. Nonetheless he maintains that the intervention of the agent-causal capacity (however causally structured it may be) as the ultimate uncaused arbiter of any internal disputes satisfactorily counters any possible regress, presumably also restoring UR and ensuring O'Connor's agents have the Gold Standard freedom of the will. *Contra* Strawson, then, agents can indeed be *causa sui* – the cause on oneself. Ultimately, though, O'Connor is forced to acknowledge that freedom likely "comes in degrees," with our options constrained by a CEP formed from past experiences, and the relative probabilities for choosing between them affected by it; which will impact on our freedom despite the presence of the agent-causal capacity. This might seem a broadly sensible approach, but what it really shows is the difficult (if not impossible) task libertarian theorists have set themselves by attempting to devise theories that meet the Gold Standard and circumvent the *Catch-22 of Libertarianism*. O'Connor, like Kane and Ginet, is trying hard to avoid determinism and arbitrariness, but really (again like Kane and Ginet) seems only to be producing a model in which his agents are subject to both.

2.7.5 The *Criteria*

O'Connor's agent-causal theory focusses on substance causation, non-Humean event causation, irreducible agents, emergent properties, and (what some have referred to as) the God-like property of agents to be *prime movers unmoved*; all of which could arguably make the approach appear less instinctively acceptable. Whilst this is not necessarily a reason to take a view less seriously, it does perhaps place the burden of responsibility on its supporters to explain why it is worth accepting this apparently quite different view of nature and the agent’s place and function within it. O’Connor acknowledges that some will find his ontological commitments hard to accept and thus goes to great lengths to explore the metaphysics of the approach (as well as the benefits of subscribing to the agent-causal model) in order to argue why such commitments are acceptable. The question now is whether, in the process of doing this, he has managed to develop a theory which can successfully meet my *Criteria for Coherence*.

As before, the first three criteria are broadly standard for all libertarian theories so can be taken together.

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.
P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.
P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.

As with the other two philosophical positions considered, O’Connor’s success in meeting the first three criteria would seem mostly assured, though (again as with the others) not without certain caveats. O’Connor’s agent-causal theory, whilst generally an indeterministic theory, does present a process for decision-making which could arguably be said to be “sandwiched” by determinism. Prior to any decision, we have varying objective propensities causally structuring the agent-causal capacity which presumably stem from internal motivations that arise deterministically via the CEP (if they are not to be labelled arbitrary). Post decision, we have the agent-causal capacity itself causally determining the ultimate action. It is the middle ground where its indeterminist credentials are formed; in the well-defined causal gap between the two. With the agent-causal capacity (by definition) not able to be the subject of further causation by anything, the indeterminism is restored. The agent, therefore, would indeed seem free to act or act otherwise whatever the past circumstances but it could be argued that this is somewhat attenuated by antecedent, structurally-causal propensities which O’Connor claims influence the agent-causal capacity to a greater or lesser degree. However, despite this, I would argue that the first three principles in the *Criteria* have broadly been met.

P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.

Prima Facie, O'Connor's model would seem the strongest of the three libertarian approaches in this area as he presents the agent-causal capacity as something which (again, by definition) cannot be either random or determined by prior circumstances, and yet is itself *fully determining* – seemingly offering a level of agent-control superior to the other theories considered so far. On closer inspection, however, it seems possible to argue that the theory is, on the contrary, more just a mixture of *both* determinism and randomness. As just discussed, if the causally structuring process is not to be entirely random then it is presumably governed by prior motivations generated deterministically from the agent's CEP, meaning that the limited options available to the agent for consideration arise (and are assigned their relative strengths) in a deterministic fashion. What follows this process is an entirely uncaused-causing as the agent-causal capacity steps in to resolve the situation. Thus, either we allow that the agent-causal capacity is all-dominant and (due to its entirely uncaused nature) admit to apparent arbitrariness in the process, or we give more influence to the causally structuring phase, which makes the process appear essentially deterministic. Either way, the principle has not been met.

P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.

Whilst it might be possible to argue that – more than any other theory considered thus far – the agent here is clearly (at least is a key sense) the "cause of herself," the regress really just seems to shift to before the agent-causal capacity is utilised, to the causally structuring phase. As O'Connor acknowledges, the requirement for deliberation along with the prescribed options and the varying strength of the accompanying propensities all arise essentially outside of the agent's control and are thus all potentially subject to an inevitable regress. Yes, the agent-causal capacity then steps in and (completely uncaused) causes whatever outcome it desires, but at the cost of an apparent and significant disconnection from what has gone before (in a similar way to Ginet's basic mental action).

Thus, even if we do allow the agent-causal capacity to fully escape the regress affecting the prior process of structural causation we would be left with a capacity that appears to act entirely arbitrarily.

P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

Of all the philosophical theories considered thus far, the agent-causal approach is arguably asking the most when it comes to accepting what could be described as unusual ontological requirements. The very existence of the agent-causal capacity, as O'Connor presents it, let alone its ability to operate and interact within a biological organism – and do so alongside other, more standard (albeit non-Humean), event-causal processes – appears arguably somewhat mysterious and obscure. Although O'Connor, it is fair to say, goes to great lengths to outline the metaphysics of his approach and argue why they should be accepted, there is still no real detail for precisely what such a faculty actually does or how it might operate. As Daniel Dennett has put it:

> Agent causation is a frankly mysterious doctrine, positing something unparalleled by anything we discover in the causal processes of chemical reactions, nuclear fission and fusion, magnetic attraction, hurricanes, volcanos, or such biological processes as metabolism, growth, immune reactions, and photosynthesis. (Dennett 2003, p.100)

As already stated, whilst a theory seeming to be somewhat original is certainly not itself a reason to dismiss it, if it also is not apparently answering any of the objections levelled at its group then there seems no benefit in accepting the cost of such ontological complications. Despite O'Connor's best efforts to normalise the approach, the theory does not meet principle six of the *Criteria*.

2.7.6 Conclusion

Whilst an impressive attempt to spell out the metaphysics of the approach and present the agent-causal theory as a genuinely competitive alternative model, the ontological commitments of the theory remain difficult to accept and, more importantly, accepting them (and the model based on them) does not satisfactorily answer the key objections I have been considering or meet the criteria I have developed. The agent-causal capacity itself is somewhat mysterious, as are its origins, operations, and interactions.

Additionally, despite setting out first and foremost to restore sufficient agential control – argued by O'Connor to be lacking in all other libertarian approaches – the increased focus on structuring propensities of presumably deterministic origin, which is combined with an agent-causal capacity apparently disconnected from them, make any claim to an enhanced control seem dubious. In the end, it could be argued, it is O'Connor's dedication to his task of making the theory more acceptable that is his undoing. In moderating the agent-causal view and incorporating event-causal and non-causal facets, he is actually chipping away at some of the potentially positive aspects of the agent-causal theory (such as the unapologetic vision of the agent as *prime mover unmoved* and the sole true arbiter of her fate) without improving its position against the other theories.

Ultimately, a theory based on non-standard causal foundations and requiring uncaused-causation by enduring, irreducible substances, emergent phenomena, and almost God-like powers of self-creation – yet still not apparently solving any of the key issues facing libertarian theories – is always going to be a difficult approach to sell. Richard Taylor, himself a proponent of agent-causation, sums the position up well:

> One can hardly affirm such a theory of agency with complete comfort…and wholly without embarrassment, for the conception of men and their powers which is involved in it, is strange indeed, if not positively mysterious. (Taylor 1974, p.58)

2.8 Where to Next?

In Part One of this thesis, I have considered key theories from the three main libertarian positions and evaluated each against my *Criteria for Coherence* – which I have argued represents the stated goals libertarians have set for themselves and which is designed to test whether the models for free action considered manage to achieve all such goals in a coherent fashion. I have argued that none of the theories considered has satisfactorily met the principles contained within the *Criteria* and thus cannot, by these standards, be considered coherent.

Robert Kane, in his event-causal approach, attempts to remove both the problematic regress and any arbitrariness from his model of free will with the introduction of his regress-stopping, chance-reducing self-forming action's (SFAs) and by imposing conditions that such free actions are required to meet; including successfully choosing one or another option (despite a probability of failing to choose it and thus doing something else), successfully doing what one *wanted* to do (even if what one "wanted to do" encompasses more than one course of action), and psychologically/phenomenologically endorsing the choice as *one's own* and recognising the choice as something one wanted to occur. However, for all its positives, the theory ultimately fails to meet the criteria I have set down. This is due, predominantly, to a seemingly inescapable level of arbitrariness inherent in Kane's model (which is generated by the attempts to see off any regress) *and* some apparent residual regressive complications (which are brought about by trying to reduce the arbitrariness); both of which (arguably) could be considered unacceptably control-reducing (and thus responsibility-reducing) factors.

Carl Ginet, with his non-causal approach, attempts to break the regress by arguing that free actions need not be caused by anything at all, including the antecedent states of the agent (such as beliefs, desires and so forth). Actions, Ginet argues, start with a basic, causally simple mental event (such as a volition), and what makes such mental events an action at all is the accompanying "actish" quality. It is this phenomenal quality that makes the agent the subject of her action, and then what makes such an action free, is the fact it is not causally necessitated by antecedent events. Once again, it is difficult to see how Ginet's theory fully avoids *either* determinism or arbitrariness and it carries the additional detraction of being arguably quite obscure in general, with a non-causal picture that is difficult to comprehend and with no real positive account offered for the ontological relationship between an agent's CEP and her actions.

Timothy O'Connor, with his agent-causal approach, seeks to bestow upon the agent an absolute form of control that he believes is missing from both the event-causal and non-causal theories. The agent on his account – as an enduring substance irreducible to events – has full, free, and direct control over their decisions and actions by way of a special, primitive causal relation (also not reducible to causation by events), which allows them to directly cause their actions or intentions to act. Yet again, however, it is difficult to see how O'Connor can claim to have escaped either determinism or arbitrariness and, in addition, his theory entails arguably unusual ontological commitments which there seems to be no real benefit in accepting.

With none of the theories discussed up to this point successfully meeting the *Criteria*, and thus not managing to offer a coherent model that fully achieves the Gold Standard free will libertarians have set out to attain, it would appear we perhaps need to look outside of the main three libertarian camps for an alternative indeterminist approach that might fare better. A number of theories have been proposed in recent years by researchers working in other branches of academia (predominantly the sciences and related fields) – inspired by developments in neuroscience and fundamental physics – which claim to add new perspectives on the issue of freedom and responsibility, and which I believe are well worth considering to assess whether they have made any headway towards a more satisfactory meeting of the main libertarian ideals.

Part Two

The Scientists

3.1 Introduction

The question posed at the beginning of this thesis is as follows: Is libertarian free will an inescapably incoherent concept? I have argued in Part One that the theories considered (one from each of the three main strands of libertarian thought) fail to convincingly demonstrate that the libertarian position can be considered coherent and that this is mostly due to the inability of the proposed indeterminist models to fully meet the principles laid down in my *Criteria for Coherence* or escape the apparently inexorable luck/regress *Catch-22 of Libertarianism*. Philosophy, however, is not the only academic arena in which debates concerning free will and responsibility are being played out and such discussions are becoming more and more popular in research areas such as psychology, neuroscience, and physics, where new theories are being produced that warrant close attention.

Researchers in the sciences and related fields have, of course, often turned their attention to philosophical topics; with those within the scope of Philosophy of Mind (such as consciousness, intentionality, perception, thought, and models for free action) always being of particular interest. Given the obvious importance of the relevant physical processes in such debates, it is not surprising that philosophers have also regularly drawn on supporting scientific evidence to add credibility to their theories as well as imbue them with the rigour often seen as exemplified most in the sciences. Robert Kane, whose work I considered in detail in Part One, often discusses the need for libertarian theories to align themselves satisfactorily with scientific thinking[45] and Timothy O'Connor also goes to significant lengths in his attempts to demonstrate that his emergent agent-causal capacity *can* be grounded in what he considers to be majority scientific thinking.[46]

As the physical brain is a (if not *the*) key biological organism involved in human action and decision-making, interest from neuroscience, biology, chemistry, and so forth, is inevitable (not to mention the social and cultural significance of such issues as freedom and responsibility) and such debates have also been spurred on by famous experiments in neuroscience and psychology.[47] And with the development of quantum mechanics and its apparently indeterministic foundations, interest has also developed in physics and mathematics as

[45] See Kane 2005, Chapter 5 and 1996, p.212 as examples.

[46] See O'Connor 2000, chapter 6.

[47] One particular example being the experiments conducted by Benjamin Libet and his followers, which focussed on recording electrical activity in the brain of their subjects whilst also asking them questions concerning their conscious intentions. The results have been used to argue that conscious perceptions occur, in fact, only *after* a decision to act has been taken physiologically – too late for conscious causal efficacy – and, as such, have cast doubt on the possibility of genuinely willed action (except perhaps the capability to consciously reverse or overrule a non-conscious impulse to act). More on this is section 3.3 below.

researchers have sought to incorporate this indeterminism – as well as other features such as uncertainty, probability, superpositions, observer participation, and non-locality – into innovative new theories of human action and free will. As there is now a large and growing body of work from such disciplines focusing directly on the philosophical issues under discussion it seems only sensible that I also consider some of the most relevant models to see how they fare against the *Criteria for Coherence* that I have developed.

3.2 Peter Tse's Criterial Causation

In an article in *New Scientist* (Tse 2013a) neuroscientist Peter Ulric Tse claims to have identified a particular brain mechanism which he believes means free actions *can* navigate the seemingly elusive "middle path" between the unfree extremes of deterministic inevitability and indeterministic randomness. If proven correct, it seems he is potentially offering a way to achieve the libertarian Gold Standard freedom of the will as well as a way *out* of the *Catch-22 of Libertarianism*. According to Tse, the reason many researchers (in both science and philosophy) have been driven towards viewing indeterminist free will as an illusion is down to a misapprehension of how neurons in our brains encode and transmit information. The traditional picture of neuronal activity as solely a succession of action potentials (also known as *spikes*) cascading through neural circuits – with one spike able only to trigger another in a cause-and-effect, feedforward, linear fashion – gives the impression of our brains as deterministic biophysical machines. Tse argues, however, that this image does not represent the whole story and such spikes can actually alter how other neurons in the circuit will respond to future inputs without necessarily triggering them immediately – like "changing the combination on a padlock without opening it." (Tse 2013a, p.28)

On Tse's model, our brains respond to situations requiring decisions or actions by setting up relevant criteria in neurons and neural circuits, which will then need to be met by subsequent inputs in order for the neurons to fire. If we then combine the brain's use of this mechanism to set up criteria for future firing (which Tse has termed "Criterial Causation") with neuronal inputs arriving at the synapses that are fundamentally indeterministic in nature due to the amplification of events at the quantum level – and which will either cause the neurons to fire or not (perhaps leading to future firing and/or further re-setting of criteria) – *and* have it all playing out within the arena of an active and coordinating consciousness, we have a model for human decision-making and action which realises a particular form of top-down mental causation that is not completely deterministic nor completely random: thus meeting both the Alternative Possibilities (AP) condition and the Ultimate Responsibility (UR) condition and delivering the Gold Standard in a coherent form.

> [C]riterial causation offers a path toward a strong free will that passes between the unfree "Scylla" of determinism and the equally unfree "Charybdis" of randomness. (Tse 2013 p.22)

In his main text on the subject, *The Neural Basis of Free Will*, Tse makes it clear that his purpose is to debunk the notion that – due it being apparently underpinned by entirely deterministic and mechanical neural processes – there can be no genuine mental causation (and thus no causally efficacious *will*) and to propose a model that *does* offer an agent sufficient control over her actions. He outlines his own conditions which must be met in order for an agent to realise what he terms a "strong free will":

> We must have (a) multiple courses of physical or mental behavior open to us; (b) we must really be able to choose among them; (c) we must be or must have been able to have chosen otherwise once we have chosen a course of behavior; and (d) the choice must not be dictated by randomness alone, but by us. (Tse 2013, p.133)

These conditions, it will be noted, align closely with my own *Criteria for Coherence* and thus Tse's intention is to hold his libertarian theory up to a similar standard. To achieve this strong free will, he proposes a three-stage model of mental causation and free will.

> [A]ccording to which (1) new physical/informational[48] criteria are set in a neuronal circuit on the basis of preceding physical/mental processing at time t1, in part via a mechanism of rapid synaptic resetting that effectively changes the inputs to a postsynaptic neuron. These changes can be driven either volitionally or nonvolitionally, depending on the neural circuitry involved. (2) At time t2, inherently variable inputs arrive at the postsynaptic neuron, and (3) at time t3 physical/informational criteria are met or not met, leading to postsynaptic neural firing or not. Randomness can play a role in the first two stages, but not substantially in the third, because intracellular potential either passes the threshold for firing or it does not. Such criterial causation is important because it allows neurons to alter the physical realization of future mental events in a way that escapes the problem of self-causation of the mental upon the physical. (Tse 2013, p.148)

[48] It should be noted that the terms 'information,' 'informational' and so forth are used frequently throughout Tse's work (and indeed the work of a number of neuroscientists) without any real explanation as to what, precisely, is meant by them. Information at times seems to describe simply the collective activity of neurons, or the brute content of what is being conveyed to different parts of the brain. At other times it appears synonymous with mental activity itself. At others it seems to be the propositional content of neural activity. Tse speaks of information or patterns of information (generated by the activity of vast neuronal circuits) being genuinely causal over and above the individual physical events. Ultimately usage of the terms is unclear and at times confused, which is unfortunate as they appear central to Tse's theory. He does seem to acknowledge this problem as he himself asks the question at one point: "what do we mean by "information"" (Tse 2013, p.2) but then does not go on to answer it.

Tse is ultimately trying to achieve two main goals. Firstly, he wishes to present a novel model of mental causation that is top-down and non-reductive – but still efficacious and non-random – and will do so by focusing on causation by *patterns* of energy, which he states are informational (and possibly also mental/conscious), instead of mere *amounts* of energy. Secondly, he wishes to use this model of mental causation to construct a successful libertarian theory of free will that avoids the problems raised by the Basic Argument and the Luck Objection. In the former case, by having mental events change the neuronal basis for *future* – not present – physical events, and, in the latter case, by removing the amount of influence random factors can have over human decision making and action.

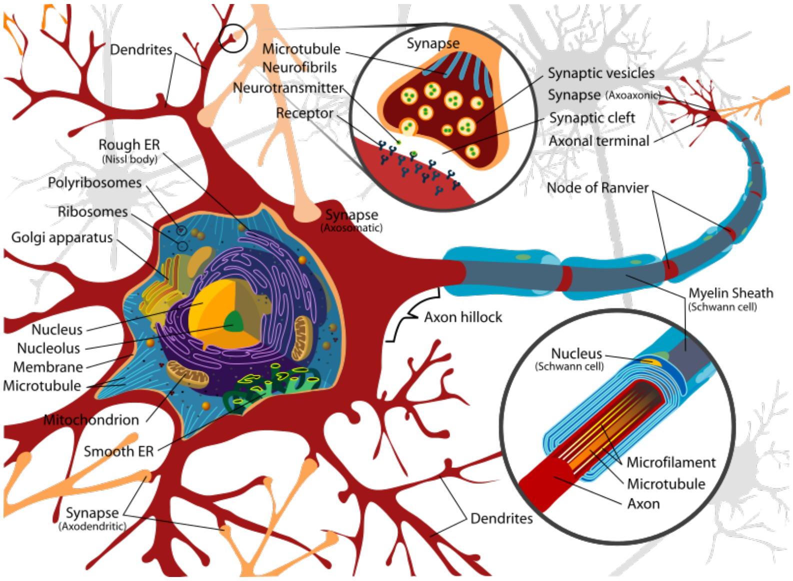
There is a great deal to unpack here in order to properly assess Tse's approach. I shall begin with a consideration of the neural mechanisms he is proposing, before moving on to consider his model for Criterial Causation, the role of indeterminism, the role of consciousness and how it all might tie together to form a model for a top-down mental causation. I will then consider whether his theory makes any further advances against the Luck Objection and/or Basic Argument before finally assessing it against my own *Criteria for Coherence*.

3.2.1 The Biology of Free will

The human brain is likely comprised of somewhere in the vicinity of 80 to 100 billion neurons, with each neuron having potentially thousands of dendrites forming connections with tens of thousands of other cells. The numbers involved are clearly astronomical and our understanding of exactly how the neural machinery of human brains works is fittingly limited, however, neuroscience has progressed rapidly over the last few decades and there are now numerous theories concerning the inner workings of neurons and synapses. A neuron up close appears a rather blunt instrument; an organic switch operating mindlessly in connection with linear sequences of electrical and chemical signals. The activity of each neuron at first glance appears wholly (and dumbly) reliant on the inputs from those that precede it, creating a mechanical and seemingly deterministic picture of human action reaching causally back into the past. There are no immediately obvious causal gaps which a libertarian theorist of free will could perhaps exploit, however Tse believes suitable gaps *do* in fact arise, generated by the way in which synaptic connections can be rapidly altered in a coordinated fashion by neurons themselves.

The Structure and Function of the Neuron

Neurons are the main building blocks of the brain and central nervous system and are Tse's key to unlocking the self-generating power of the brain. Though heavily outnumbered by their supporting glial cells, neurons are believed to take on the key role of computation and information processing and use electrical impulses and chemical signals to transmit information. Though there are various different types of neurons with different roles operating in different areas of the brain and spinal column, for our purposes a focus on the standard representation will suffice.



(credit Wikipedia)

The traditional view of the working of neurons does indeed seem to depict the cell as a brute, natural, all or nothing, feedforward phenomena. As a hugely overly simplistic description:

1. The dendrites surrounding the soma receive neurotransmitters via synapses with other neurons, which cause electrical changes that are then assessed by the soma at the axon hillock.
2. If the required threshold is met (a strong enough signal coming from the dendrites) an action potential travels down the axon to the axon terminals.
3. The action potential causes the release of neurotransmitters at the axon terminals which diffuse across the synaptic cleft to the dendrites of the next neuron(s) in the chain.

The picture painted of the structure and function of the brain by this basic description of its key component does indeed appear a wholly mechanistic one. Whether deterministic or indeterministic, if we take the brain neuron as the basic substrate of human decision-making and action, it appears simply a cause-and-effect process of neurotransmitter release, action potential, neurotransmitter release, and so on.[49] The firing of each postsynaptic neuron appears dependent on the strength or type of the input it receives from the previous neuron or neurons, creating a web of linear transmission across the neural circuits in the brain. Individual neurons on this view appear to act independently of any higher-order mental influence and thus on the often-popular reductionist picture (particularly favoured within the scientific community), any kind of higher-level conscious mental events wholly reducible to neuronal activity would seem epiphenomenal and inefficacious.

Based on such a picture, we might imagine that Jessica's brain, on seeing the elderly gentleman fall, would react entirely mechanistically in response to the stimuli received from her senses, producing an automatic cascade of neuronal activity. Sensory neurons from the photoreceptor cells in her retinas would convey information about what she is seeing to the relevant parts of her brain for processing (as would others from such sources as the hair cells of her ears and the olfactory cells in her nose) and this information would be transmitted through the neural circuits (largely via interneurons), from neuron to neuron, as electrical and/or chemical signals, with neurons either firing or not firing dependent only on the strength of the inputs, until ultimately (perhaps after some kind of neural web-based assessment in a central processing area representing her current established psychology, or CEP) the motor neurons become involved, which signal contractions of the relevant muscles for her to either drive away or exit her vehicle.

Tse's point is that any suggestion on this picture that higher-order mental events associated with what we would likely label as conscious reasoning – which he believes to be wholly realised in physical events – could alter the very physical events on which they are based would be unacceptably circular. Indeed, he argues, if we accept this traditional view of neuronal workings, we must accept that the driving force behind all decisions and subsequent behaviour is wholly the mechanistic action of the stimulus-response, feedforward, all-or-nothing neurons and any kind of higher-order mental causation of the sort Tse believes would be required for free will is simply ruled out.

[49] Or electrical stimulus in the case of electrical synapses at gap junctions; the principle is the same.

Tse's Alternative Picture

Tse's solution to this problem, in the first instance, is to present an alternative picture of neuronal activity in which neurons, rather than solely responding crudely to stimuli from other neurons, can actually dynamically alter their synaptic properties and those of the neurons around them (extremely rapidly if necessary) to pre-set criteria that will determine whether or not an action potential is generated when an input next arrives at the synapse. Tse is attempting to change the perception of neurons as solely feed-forward linear classifiers – largely passive receptors – whose firing is decided exclusively by whether or not the summation of the electrical potential of the incoming stimuli passes a fixed threshold at the axon hillock; presenting them more as active players in cognitive processing.

The scientific explanation for exactly how our neurons might achieve this (biologically speaking) is expectedly complicated[50] but, in brief, the changeable properties of synapses that Tse cites as being key are the "weights, gains, and temporal integration properties" (Tse 2013, p.22), which can be altered to recode the inputs that will make a neuron fire in the future. These changes can be made without triggering an action potential in the affected neurons and can happen in a variety of different ways. They can occur over an extended period of time or extremely rapidly (within milliseconds under the right conditions); the latter of which, Tse argues, is essential if they are to be relevant for conscious decision-making. Due to the connectivity of neurons in neural circuits, a change to only one neuron can have a major effect on the entire circuit.[51]

> The physical criteria for neuronal firing change through operations like excitation, inhibition, hyperpolarization, subthreshold depolarization, long-term potentiation, the post-action potential refractory period, the dynamic changing of synaptic weights on existing inputs, growing of new dendrites or synapses, expression of proteins in the membrane, changing of thresholds, apoptosis, and changes in the resting potential. (Tse 2013, p.74)

Of particular importance are the NMDA (N-methyl-D-aspartate) receptors, found mostly in dendritic spines, which are believed to have a key role in controlling synaptic plasticity due to the way they permit

[50] Tse discusses it in great detail in chapters 4 and 5 of his book.

[51] In his *New Scientist* article Tse likens this change to neural circuits to railway switches: "Just as railway switches must be flipped to allow trains to pass, synaptic weights must be reset before brain signals can follow one path through a neural circuit instead of other possible paths." (Tse 2013a, p.28)

a rapid influx of calcium ions across the cell membrane. These proteins are worth mentioning because they also play an important role, according to Tse, in introducing the much-needed randomness to the macroscopic level in the form of spike-timing uncertainty, as well as transitioning neural networks into a "bursting mode" which underpins attentional binding and is essential for criterial causation and the ultra-rapid synaptic resetting. [52]

Tse acknowledges that (as with much of the workings of the brain) *exactly* how the rapid changes on the weights of the inputs at the synapses occurs is not yet well understood and his model is largely theoretical, but the role of such processes is key to understanding information processing in the brain; just as much as the trains of action potentials themselves. It is the manipulation of the action potentials by the criteria setup at the synapse, which affects the inputs, that really imparts the information into the circuit. Not the brute all-or-nothing activity of the action potentials. According to Tse, it is important to know the information that caused the neuron to fire, not just know that it has fired: "[O]bserving an action potential without knowing about how it is "filtered" by synaptic weighting cannot tell you what information it conveys, if any." (Tse 2013, p.69)[53]

The point, for Tse, of introducing such an alternative model of neuronal activity is that, if neurons can indeed effectively rewire themselves and each other (and thus entire circuits of which they are a part) in such a fashion, then by extrapolating upwards he can argue that our thoughts, actions, and decisions – which he believes to be ultimately fully realised in the activity of our neurons – are not solely at the mercy of ballistic sequences of cascading action potentials. Instead, Tse argues, such mental processes can also exert a form of top-down causation by way of criterial recoding which can temporarily (or perhaps also permanently) change the degree to which synapses trigger future action potentials, and thus actively (and presumably deliberately) encode new criteria for future firing. Whilst it seems as though there may be evidence that such a neural mechanism for altering the synaptic properties of neurons is a possible (and perhaps even likely) candidate for how our brains operate at the neuronal level, the important question for our current discussion is whether an acceptance of this biological model as a foundation for a libertarian theory of free will (which is what Tse is ultimately attempting to present) would lead to a theory that fares any better than those already considered when assessed

[52] See Tse 2013, Chapter 5 for more on this.

[53] One interesting example Tse uses to illustrate his point is that of a voice. The shape of the mouth (the encoded synaptic criteria) is more important that the air (the brute action potential) that is subsequently pushed through. See Tse 2013, p.69.

against the *Criteria*. Even were the model proven to be a correct picture for how our brains work, could such potential for control over our brain biology afford an agent any more control over her decisions and actions than the brute external-stimuli-to-internal-spike-train-to-external-action in such a way as to tackle the *Catch-22* of such a process being subject to either randomness or an inevitable regress?

Though providing interesting scientific detail on the potential neural mechanisms involved in human action and decision-making, it is not immediately clear how these modifications in any way bolster Tse's claim that he has achieved the Gold Standard libertarian free will; as judged by the criteria I (and Tse himself) have laid down. By moving the locus of control from within a wholly reductive picture of algorithmically firing neurons – unaffected, it seems Tse is suggesting, by any coordinated workings of a higher-order consciousness or self – to a higher-level processing capability that can deliberately coordinate such firings by synaptic manipulation does not seem to remove the circularity Tse is hoping to address. Any problems I have argued are inherent in libertarian theories would seemingly be just as problematic for the higher-level processors, in a similar way that they remain a problem as much for the formation of the antecedent desires as they do for the enacting of the choices they guide – as discussed on a number of occasions in Part One. Tse, however, argues that acceptance of such a model – when combined with other key factors such as indeterminate synaptic inputs and a coordinating consciousness – *does* add something to the libertarian picture and can produce a model that *would* therefore meet the Gold Standard.

3.2.2 Criterial Causation, Indeterminism and Consciousness

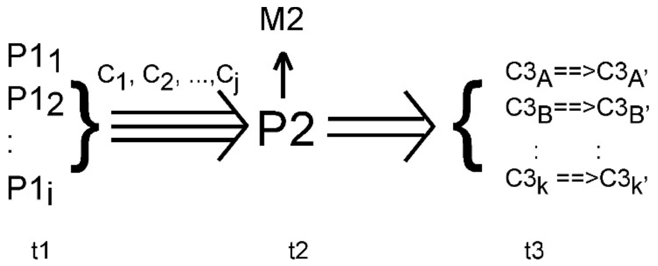
Tse's complete theory for a neuronally-based, indeterminist free will has three main parts to it:

1. The neurons in our brain and central nervous system have the ability to rapidly rewire each other (and themselves), which changes the degree to which synapses trigger future spikes and thus imposes causal criteria for future activity – hence the term, Criterial Causation.
2. Neuronal inputs are inherently variable and unpredictable due to microscopic indeterminacies in the synapse which, in conjunction with criterial causation, prevents solely determined outcomes.
3. Internally manipulable conscious thoughts are required in combination with 1 and 2 in order to make our decisions and actions truly free and make us fully responsible for them.

I will consider each of these in turn.

Criterial Causation

Tse presents the following schematic of his criterial causation model:



Tse's diagram for Criterial Causation (Tse 2013, figure 3.1 p.26)

Inputs arriving via the synapses from a variety of neurons (P1.1, P1.2 etc) at the dendrites of the target neuron are then assessed in the soma against whatever criteria (C1, C2, C3 etc) has already been programmed into the neuron in previous interactions.[54] This criterial assessment (the crux of criterial causation, as represented by the triple arrows) then decides whether the neuron fires at t2 and P2 is released. If it does fire, the resulting action potential is not just confined to a binary role of either triggering or not triggering subsequent spikes in the adjoining neurons. It is also capable (most likely by rapid bursting) of acting as a modifier as well as a trigger, changing the synaptic weights on the post-synaptic neurons and thus altering their currently encoded criteria for firing, without necessarily causing them to fire at t3. Different inputs may now make the post synaptic neurons fire than previously (e.g. criterion C3A changes to C3A') and Tse would want to say that potentially the 'informational' criteria

[54] Tse defines "criteria" as: "a set of conditions on input that can be met in multiple ways and to different degrees" (2013 p.22). It would appear that there are in fact a multitude of different combinations of conditions that could be imposed – as well as many variations in the individual conditions themselves and also the possibility of dynamic altering of conditions – resulting in a generally wide variability in the criteria available to neurons to employ. Unlike rules, criteria can be assessed based on the degree to which they have been met and neurons employing the criteria could act on an aggregate of criterial fulfilment. All this affords neurons a great deal of flexibility in how they create and employ criteria, which is presumably what Tse wishes in order to depict a less rigid and mechanical picture of neuronal function. He in fact states that the malleability and trade-offs required for criterial assessment is what gives decision-making its deliberative rather than ballistic character: "Criteria are generally not so specific that they cannot be met by a range of inputs. Indeed, criteria can generally be satisfied in multiple ways". (Tse 2013, p.74)

required to make them fire has also changed.[55] The change from the triple arrow prior to t2, to the double arrow after t2, represents the move to *non*-criterial causation as the assessment has presumably already concluded and the action potential (or rapid succession of action potentials) is either on its way or not, ready to either trigger the adjacent neurons or recode their criteria for firing in the future.

As mentioned above, an important part of Tse's model is that such criterial encoding means that *patterns* of energy, rather than mere *amounts* of energy, can be causal. That is, patterns of neural activity which realise information can be causal over and above the physical transfer of energy or particle collisions – but only if there is a biological mechanism in place that is capable of recognising and responding to such patterns, such as his criterial causation.

> A central message of this book is that *patterns in input can be genuinely causal only if there are physical detectors, such as neurons, that respond to patterns in input and then change the physical system in which they reside if the criteria for the presence of a pattern in inputs have been met*. Neurons that respond to patterns in input, such as temporal coincidences, that carry information about spatial and other patterns in the world, such as, say, the spatial pattern of the constellation Orion, can be thought to realize a downwardly causal physical process…[P]attern detection among neurons cannot be reduced to the localistic transfer of energy among colliding particles. A neuron does not absorb or transfer energy like a billiard ball. It assesses its inputs for satisfaction of physical/informational criteria, which, if met, lead to output. These criteria do not assess the amount of energy in inputs. They assess patterns in input. (Tse 2013, p.9)

This is important because Tse is identifying such patterns of neural activity with mental events and it is such mental activity that he wishes to show should not be considered solely epiphenomenal; thus aiding his argument that genuine mental causation is a real possibility. He is arguing that mental events *being realised* in physical events does not mean that they *must* be epiphenomenal, because such events (as patterns of neural activity), in conjunction with criterially infused neurons and neural circuits, can be causal in a way that is not wholly reducible to the transfer of energy. This, he believes, allows us to talk about causation at the level of information (mental) processing in the brain rather than at the level of elementary particles.

[55] In the *New Scientist* Article Tse gives the following example: "This means that a neuron could now be driven by an input that, moments before, might have contributed nothing to its firing. For example, a nerve cell that has just responded to a touch on your forehead could now respond to someone stroking your hand." (Tse 2013a, p.28)

The problem is that the way Tse has represented the involved mental state or states (M2) in his schematic seems to contradict its portrayal as causally efficacious. Tse wishes to say that when P2 is released it realises the information, M2, and he states that the single arrow here represents "the supervenience relationship of the mental on the physical" (ibid., p.26), which indicates that he sees the information realised in this process as essentially a mental event. The addition of the mental element to the schematic is necessary, as part of his overarching goal is to argue that his criterial causation model is ultimately a model of top-down mental causation which evades the causal exclusion arguments made against non-reductive physicalism.[56] But it remains unclear from Tse's arguments exactly *why* such a mental event should not just be considered epiphenomenal. Tse is not arguing any identity between P2 and M2, but that M2 supervenes on P2, and thus there seems to be no obvious reason to assume the mental event (on this diagrammatic representation at least) is doing any work at all – particularly as the arrow it sits above is only pointing in the one direction: from the physical P2 up to the mental M2. Perhaps Tse would wish to argue that the mental event is not simply supervening on P2 alone, but on countless concurrent incidences of similar occurrences across a neural network,[57] and this representation is deliberately simplified in order to communicate his ideas effectively; with the real mechanism likely involving millions of P2s with millions of arrows all pointing towards a single M2.[58] M states would then form part of a web of interactions and patterns in which they could have a form of causal efficacy. But it could always be argued by those that view mental states as epiphenomenal that if every M state is wholly realised in a P state (or a collection of P states) then the removal of the M states from the diagrams would not affect any of the neural activity and subsequent human behaviour.[59]

[56] Tse, in particular, targets the arguments developed by Jaegwon Kim. See Tse 2013, p.123 and also Appendix 2, p.247.

[57] Tse indeed repeatedly comments that a single neuron does not usually convey much useful information in isolation it is the behaviour of the circuit as a whole that is most important. See Tse 2013, p.7: §1.11 as one example.

[58] We have just discussed Tse's desire to show how patterns of information (of which presumably events such as M2 form a part) can somehow be causal over and above their physical basis, but this still requires the events constituting the pattern to be physical. What is also important is that we are not just talking about one neuron firing (or not firing) after another. Criterial causation involves massive chains, circuits, or sequences of neuronal assessments of their informational inputs. This wider net, or pattern of criterial assessments, and the potential for continual and rapid alteration of the criteria, is what is important to allow for the collective information being processed by the neuron to be causal over and above the mechanics of its physical substrate.

[59] The debate on mental causation and reduction is, of course, extensive with volumes devoted to it and I do not propose to delve too much into it here.

Without some form of feedback from M2 into the circuit or another arrow from M2 aimed at a different circuit, it still remains unclear what causal effect M2 could have on the system.

In addition to this, Tse is a committed physicalist and he is clear that he considers mental events just *are* physical events, adding further questions as to what exactly he believes M2 *is*? What is its status, ontologically speaking? Tse appears to suggest that M2 is a physical event, but one of a different kind to P2. He uses the term "physical/informational" to describe such events and discriminates them from other states by stating that all informational events are physical events but not all physical events are *also* informational (in fact only very few, which are organised in such a way that occurs when criterial causation is in play, will be).

> Multiple physical causal chains are possible at any moment, assuming indeterminism. Only a minute subset of these will also be informational causal chains. Criterial causation is powerful because it allows only that subset of possible physical causal chains that are also informational causal chains to occur. In cases exhibiting criterial causation, informational causal chains just are physical causal chains, and vice versa. For physical systems that realize criterial causation, to say "such-and-such physical events caused such-and-such physical events to happen" becomes a different way of saying "such-and-such information caused such and such information to happen." (Tse 2013, p.28)

But precisely how these particular sorts of (informational) physical events differ from more standard ones and how they would contribute to his model of mental causation over and above their more run of the mill counterparts is unclear.[60]

One argument Tse does give for the emergent type of causation he appears to favour is based on the fact that his criterial decoders exhibit a form of *multiple realisability.* He states that as the same mental state can be realised on the basis of an array of different inputs (and thus different antecedent physical states) and because criteria at each stage in a chain can differ, the information can be transformed, and thus the information can be causal. Tse also argues that patterns are not subject to the same physical laws that constrain amounts of energy because, unlike their physical basis, they lack mass and other

[60] This also leads us back to my earlier footnoted remarks that not knowing exactly what Tse (and, in fairness, many other of his neuroscientific colleagues) mean by "information" in the context of neuronal activity is arguably hampering the debate.

material attributes and can be created and destroyed. It is this fact, he believes, that allows a higher-level form of causation to emerge; so long as there is a criterial decoder that responds to the pattern.

Ultimately, it seems that on Tse's account downward causation is realised by the deliberate manipulation of the criterial detectors. He argues that due to the nature of quantum mechanics, many of the possible paths open to the elementary particles could potentially become real but, by the action of the neurons as criterial assessors/encoders/detectors, only the paths that also form the informational causal chains are realised. It is the information that decides which possible path becomes real over others, thus downwardly causing physical events.

> In the absence of criteria placed upon spatiotemporal patterns in input, criterialess physical causation is bottom-up and localistic. But by changing the criteria that future inputs must meet to make a postsynaptic neuron fire, a type of downward causation is physically implemented that biases which possible particle paths will be realized. Although physical causation is indeed closed, in the sense that only possible paths allowed by quantum physics can possibly become real…nonetheless, many possible paths could become real at each moment. Which one becomes real in a functioning brain will be a possible path that contributes to meeting the physical conditions for firing by neurons when a neuron in fact fires. (Tse 2013, p.126)

Such informational causal chains, Tse argues, are mental causal chains and they can cause one outcome over another because the information carried (once decoded in the neurons) can then trigger action that would not have occurred if the informational criteria had not been met.

In summary, Tse's criterial causation is a biological mechanism that he believes can be employed to rapidly reshape the circuitry of the brain (presumably by some form of central processing/decision making module – CEP) and thus be used to implement a volitional will. As such, it is a key component of his overall model of mental causation and his theory of free will but, as the above discussion shows, much remains unclear, especially (and importantly) in regards to the ontological nature/status and causal ability of his mental/informational events.

The Role of Indeterminism

Tse argues extensively for indeterminism as the ontological reality of our universe[61] because, although he believes the criterial causation part of his theory would, in isolation, still hold up in a deterministic universe, he does not believe genuine free will is compatible with determinism due to the view that this would result in a failure to meet the AP condition.[62] Indeterminism at the microscopic level of neuronal inputs must therefore be added to criterial causation (along with consciousness) for genuine freedom. On Tse's account, it is the inherent variability of the neuronal inputs, coupled with an agent's ability to set her own neural criteria, that underpins free will as it allows her to retain volitional control over her actions whilst still being able to "do otherwise." Criterial causation pre-sets criteria in neural circuits to guide/restrict output, but it does not *determine* the output and Tse is clear that were we to run the sequence of events over again we would not necessarily get precisely the same outcome.

Tse gives the following example as an illustration of his view:

> If I ask you to think of a politician, your brain sets the appropriate criterion in neurons involved in retrieval of information held in your memory. Perhaps Margaret Thatcher comes to mind. If it were possible to rewind the universe, you might think of Barack Obama this time, because he also meets the criterion. This process is not utterly random, because the answer had to be a politician. However, it is also not deterministic, because it could have turned out otherwise. (Tse 2013a, p.28)

Tse is therefore required to argue for a biological foundation which would allow indeterminism at the microscopic level to have an effect at the macroscopic level and, once again, it is the NMDA receptors (mentioned briefly in the previous section in regards to rapid enough synaptic re-setting) that are important in this process: "Because receptor binding is a stochastic process, it plays an important role in introducing randomness from the microscopic domain to the macroscopic domain of neural networks" (Tse 2013, p.72). The lock-and-key metaphor is often used when discussing the binding of

[61] See Tse 2013, Appendix 1, p.241.

[62] This is further evidence that Tse believes he is striving for the Gold Standard freedom of the will that my *Criteria* is designed to test against, as opposed to the more limited "free action" considered by libertarians to be all that is available to the compatibilists: "If the universe were deterministic, strong free will would be ruled out, since things could not have turned out otherwise and the weak compatibilist notion of free will would be the best we could hope for. But since physics provides evidence for ontological indeterminism (see appendix 1), a physical basis for a strong free will is in fact possible." (Tse 2013 p.134)

neurotransmitter molecules to receptors and it is often useful to view them in this fashion but, as Tse points out, they cannot necessarily be understood within the ordinary framework of classical physics. The diffusion across the synapse of the neurotransmitter, according to Tse, is actually more of a "random walk" (ibid., p.73) than a predetermined path and variability can come from both the amount of neurotransmitter released into the synaptic cleft and the varying distance needed to be traversed across the synapse.[63]

> In sum, neurotransmitter diffusion across the synaptic cleft carries both signal and noise. It is an important cause of variability in the rate and timing of neural activity, and of the neural basis of nonpredetermined but nonetheless self-selected choices...In effect, diffusion across the synapse is one of several physical mechanisms that permit the amplification of microscopic fluctuations into macroscopic variability in spike timing. (Tse 2013, p.76)

As with the above discussion on how exactly neurons can re-wire each other, the biology here is complicated and analysed extensively in Tse's book. For our purposes it seems fair to say that, again, there is evidence to suggest quantum level indeterminacies could be playing a role in the transfer and binding of neurotransmitter molecules and thus affecting the outcomes of our neural mechanics.[64] But, although Tse may view that as achieving his goal of inserting a level of non-determining volitional control into his model, the familiar question arises of how this introduction of essentially random elements would affect the sought after control. Does Tse's method of incorporating indeterminism into

[63] Tse actually argues, interestingly, that one reason for why neurons persist with the slow process of chemical transmission at synapses at all, instead of only using the faster electrical signalling at gap junctions, is just because this slower chemical process is what introduces the required variability in postsynaptic neuronal responses. See Tse 2013, p.75 for more on this.

[64] It should be pointed out that such quantum effects, however, are restricted in Tse's theory to providing the required "noise" in the system that can be harnessed as part of criterial causation and implementing genuinely novel behaviour and do not play much more of a role in cognition: "There is no need to invoke quantum nonlocality, superposition, entanglement, coherence, tunneling, quantum gravity, or any new forces to understand informational causal chains in the brain. Criteria can be realized in the input–output mechanisms of relatively large scale, high-temperature entities, such as receptors or neurons, in the absence of nonlocality effects. What is needed, however, is some degree of noise in the system that arises from amplified microscopic fluctuations that manifest themselves as randomness concerning the timing of EPSPs and IPSPs and therefore neural dynamics. Because of such noise at the synapse and within neurons themselves, there is no guarantee that identical presynaptic input will lead to identical postsynaptic output, even if time could be "rewound" and initial conditions were truly identical…It is improbable that any of the strange, nonlocal quantum coherence effects can have any influence on how neurons behave, or how consciousness or information is realized in neural events. The brain is, simply put, too "warm" to support this kind of quantum-domain coherence, and synapses are too wide to support electron tunneling." (Tse 2013, Appendix 1, p.244)

a libertarian model of free will move us any further towards tackling the key objections than the theories and models discussed in Part One? These questions will be considered in the next section, but first I will outline the third and final part of Tse's theory which he clearly hopes will address some of these issues concerning the level of randomness in the system and presumably also the requirement for some form of ultimate responsibility – the addition of an efficacious consciousness.

The Role of Consciousness

Tse's reasons for arguing that consciousness must play a key role in mental causation, and therefore in free will, is because *without* a conscious agent with intentional states (rooted in a CEP) that could put her criterial causation into action volitionally to attain goals, he believes human beings could be considered mindless zombies.

> The brain of a zombie who lacked consciousness could use this mechanism too, but we would not say it had free will. To have free will requires that our self – that which we feel directs our attention around our conscious experience – has some say in the matter of what we do or think. If consciousness plays no part in the synaptic reweighting process, there is hardly a free will worth having. (Tse 2013a, p.28)

As to the precise nature of the role consciousness plays, Tse argues it is most likely to be in providing an essential common format for relevant executive operations in the brain, such as endogenous attention, which assess potential behaviours and/or thoughts involved in fulfilling desires. In fact, endogenous attention, according to Tse, is *required* for internal deliberative processes and consciousness is *required* for endogenous attention; thus, internal deliberation can only happen at all where consciousness is in play.

> Because certain operations can only take place over conscious operands, and motor acts can follow and enact the conclusions of such operations, such mental operations play a necessary causal role in their mental and motoric consequences, and are not mere illusions of volition. (Tse 2013, p.16)

Tse is proceeding from an assumption that we have experiential states and without them we would undoubtedly act differently. Experience and consciousness, therefore, must play some key role in the causation of our actions (at both the mental and neuronal level) and this role – according to Tse – is providing a format for the assessment of various possible courses of action and allowing the internal attentional manipulation of the associated thoughts. This then leads to the neuronal recoding required for his criterial causation. Exactly why Tse believes that such operations can *only* take place consciously is not completely clear. He appears to argue somewhat by stipulation; stating there simply must be a level where competing courses of action can be assessed against the overarching goals of the agent. He thus maintains that there is an important link between endogenous attention, working memory and experience.

> Experience plays a causal role in subsequent actions by serving as the format upon which endogenous attentional and planning operations can operate in order to generate plans that can be potentiated in advance by reprogramming appropriate neural circuits such that, when the inputs that satisfy those neurons' criteria arrive, the motor act will be executed. (Tse 2013, p.225)

So, although an important part of the process, consciousness alone is not sufficiently causal. Rather, consciousness allows for the endogenous attentional operations that reach the decisions and it is these processes that then go on to impose the relevant criteria at the neuronal level. Consciousness, then (it would seem), is a form of *necessary arena* in which deliberation can occur and decisions can be made. Once these decisions are made, the framework for realising them is rapidly encoded onto neural circuits. The ultimate outcome is then generated in a partially random fashion by the arrival of variable inputs.

Whether the addition of this type of consciousness addresses any of the concerns raised by the Luck Objection and Basic Argument will be considered next but, *prima facie*, important questions about desire origination and controlled decision making seem to remain. Whilst Tse has made it clear in his work that he is developing a model to meet the Gold Standard, no real attempt seems to have been made to analyse or discuss *how or why* a particular decision is made in the first instance, which then drives the criterial encoding process, which then awaits the variable inputs. Tse's model appears only to address how such decisions could be neuronally implemented in an apparently undetermined fashion. Consciousness, as he has portrayed it, is arguably a somewhat impotent facilitator for the mechanical operations of the brain and it is not clear it adds anything substantive to what thus far seems just a biological model for the motoric enactment of a mechanistic (if potentially undetermined) process of action implementation.

Tying it all together

If we bring together all three parts of Tse's theory we are able to build the following picture of Jessica's dilemma, deliberation, and action. Upon seeing the elderly man fall, the relevant neurons from her sensory organs convey information about the situation to the cortical areas responsible for assessment, evaluation, and planning operations. Such operations likely involve endogenous attention – which necessarily requires the framework of consciousness – in order to engage in a deliberative reasoning process. Let us say that, at the culmination of such processes a decision is made to stop and help. Signals are then sent out from the initial planning regions to other brain areas involved in the next processing stage which encode into the relevant neural circuits (via synaptic re-weighting) "stop and help."[65] Once this criterial pre-setting is achieved, indeterministically variable inputs arrive at the circuit and either meet the criteria or not. If it is met, the circuit is activated in such a way as to produce an appropriate plan of action and further signals are sent out (now in a presumably determined fashion) via the motor neurons to put the chosen plan into action.

As already discussed, there are a number of problems and confusions with this picture. A key issue is that arguably the most important part of the process – the reaching of the decision to either stop and help or drive away – is left largely unexamined and unexplained and appears to occur *prior* to the enactment of the very criterial causation model Tse is proposing. A secondary problem is the level of randomness present in the theory, in regards to the indeterministically variable inputs which may or may not meet the specified criteria. Both of these issues, rather than defeating them, would appear more to *feed* both the Luck Objection and the Basic Argument.

[65] Perhaps, more specifically, Tse would wish to argue that what is actually encoded is something like "decide how to stop and help." The impetus to help is there but the variability is in how best to achieve the goal. This might fit better with his example of encoding a thought about a politician but then allowing variability for who is chosen. But if this is correct then in situations such as Jessica the key decision to help is made deterministically and the only variability is in *how* to help, which would not seem to advance the libertarian case. Part of the issue with Tse's account is he is just not clear on points of detail such as this.

3.2.3 Luck, Regress and Circularity

The Luck Objection

Tse's own conditions – which must be met in order for an agent to have free will – include there needing to be "multiple courses of physical or mental behavior open to us" and that "we must be or must have been able to have chosen otherwise once we have chosen a course of behavior" (Tse 2013, p.133). As we have seen in Part One, this AP condition that there must have been real alternative possibilities available to us at the time the decision was made (in the sense meant by libertarians) forms the foundation of the Luck Objection, which then comes to life through specific expressions such as the parallel world or rollback formulations. Whichever one we choose to focus on, the core of the argument remains the same: if we could have done otherwise, whatever the past conditions and natural laws, there is nothing within us (or indeed external to us) that has the final causal say in what we do, and thus what we do is just a matter of luck.

Tse is arguing that, despite his emphasis on the importance of indeterminism, luck or randomness is eliminated from the important parts of his model by the pre-setting of the criteria which essentially dictate what neural outputs are possible. In combining criterial causation with indeterministically variable inputs, we now have a biological mechanism for the setting up of a framework within which a desired outcome could be achieved (or at least some version of it), but which will not determine one outcome over another and which could thus have led to a different outcome. But if no particular outcome is apparently being *directly selected by the agent*, can the agent be said to actually be in control of her actions? If the active control Tse is arguing for reduces to merely the encoding of criteria in order to achieve a desired outcome – the actual fulfilment of which is then reliant on indeterministically variable inputs which may not realise such an outcome – can we truly argue such a model would meet the Gold Standard? If the variable inputs are the ultimate arbiters of our actions, does this not result in complete randomness?

In his *New Scientist* article Tse gives two examples of his criterial causation model in action. One, already mentioned, features an agent being asked to think of a politician whereupon their brain then encodes the relevant criteria "think of a politician" in its neural circuits. This ultimately leads to the agent thinking of Margaret Thatcher, but could just as easily have led to her thinking of Barack Obama instead due to the indeterministically variable nature of the inputs. The second example relates to preparing for a dinner party, in which Tse also reaffirms the necessary role of consciousness.

> Let's say you are planning a dinner party and play out various possibilities in your mind's eye. You imagine serving a steak, then realise that one guest is vegetarian, so set criteria "delicious; not meat" among synapses associated with memory retrieval. As described before, whatever comes to mind will meet these criteria yet could have turned out otherwise. Let's say spinach lasagne is the first appropriate solution that comes to mind. This solution could only have been reached through intentional manipulation of conscious thoughts, so the neural activity that gives rise to consciousness is necessary for the subsequent act of shopping for spinach. Your brain freely willed the outcome of spinach by setting up specific criteria in advance, then playing things out. Such internal deliberation is where the action is in free will, not in repetitive or automated motor acts. (Tse 2013a, p.29)

If we were to accept Tse's model, as presented in these examples, we would seem to be left with a rather limited sort of freedom at best – and seemingly not the Gold Standard we are trying to achieve. Applied to Jessica's situation, it would seem that what would be encoded into her neural circuits if she decided to assist the fallen man would be "stop and help" and then variable inputs would arrive which would dictate the nature of precisely *how* the help is given. But if variable inputs could lead to actions as different as, say, getting out and performing CPR or simply stopping the car and calling an ambulance, it seems valid to consider whether they could equally lead to her driving away altogether; apparently reinstating the problem of luck. Even if this is not a real possibility and once the instruction "stop and help" is firmly encoded into her neural circuitry the only variation is the precise nature of how that help is delivered, the key decision to help is made *before* any process of criterial causation-plus-indeterminism takes place. Similarly to the agent-causal view, then, we appear to have an agent causally implementing a desire (though backed up in Tse's model with an interesting biophysical model for how our brains might achieve this) but no explanation for how this desire itself could be formulated in a way that would not be considered either the product of deterministic neural processes or a completely random arising. As Tse is clear in his wish to avoid randomness in key areas of his model, the desire must presumably arise deterministically from Jessica's CEP, but then the precise way her desire is to be implemented is apparently decided by the random process involving variable inputs. This would, therefore, not appear to be the "middle way" Tse is claiming to have found. In reality, what we have is an initial deterministic process which is then followed immediately by an indeterministic one, and adding two things together is not the same thing as finding a middle way between them.

The Basic Argument Defeated?

Tse specifically targets Galen Strawson's Basic Argument in his work because he believes that his theory, with its focus on changing the criteria for *future* firing rather than attempting to show how mental events could alter their own physical basis in the present, is particularly well positioned to avoid the related traps of self-causation, circularity, and infinite regress (and other problems associated with the ultimate responsibility (UR) condition) that the argument highlights.

We have covered the Basic Argument in detail in Part One but, briefly, it proceeds as follows:

1. We act, in the moment, as we do due to our current established psychology, CEP.
2. If we are to be responsible for how we act in that moment we must be ultimately responsible for the formation of that CEP.
3. But we cannot be ultimately responsible for our own CEP because in order to be so we would have to have chosen to be the way we are at that moment, which would require us to choose to be a certain way at a previous moment, and so on back to a time when we were not capable of choosing anything.
4. So free will is impossible because, ultimately, we cannot be the cause of ourselves.

In sum – we cannot ultimately choose what we *do* because we cannot ultimately choose who we *are*.

Tse's strategy is to claim that Strawson's argument simply does not apply to his theory because criterial causation does not involve the apparently contradictory process of self-causation at all. Mental events, Tse argues, supervene on their physical substrates and thus for a mental event to alter the very physical substrate on which it supervenes would be tantamount to a form of self-causation and would be logically impossible. However, in his theory, mental events do not need to alter their own physical basis in order to be sufficiently efficacious. What they do is set up criteria for *future* physical activity.

> There can obviously never be a self-caused event, but criteria can be set up in advance, such that when they are met, an action automatically follows; this is an action that we will have willed to take place by virtue of having set up those particular criteria in advance. At the moment those criteria are satisfied at some unknown point in the future, leading to some action or choice, those criteria

> cannot be changed, but because criteria can be changed in advance, we are free to determine how we will behave within certain limits in the near future. Criterial causation therefore offers a path toward free will where the brain can determine how it will behave given particular types of future input. This input can be milliseconds in the future or, in some cases, even years away. (Tse 2013, p.136)

We can break down Tse's position on this as follows:

1. The Basic Argument claims that in order to be considered responsible an agent would need to somehow be *causa* sui: the cause of themselves.
2. In practice this would mean mental events altering the very physical substrates on which they are realised, which is logically contradictory.
3. In Criterial Causation, physically realised mental events enact a volitional will by altering *future* physical events (and thus physically realised mental events) by the encoding of criteria for future firing.
4. Thus, because our brains are setting criteria at T1 to be met at T2 there is no self-causation involved.
5. Our choices are in fact free because they were made by virtue of satisfying the criteria that our brain programmed into our neurons at T1.
6. We are therefore (ultimately) responsible for the characters we have and consequently for the choices we make.

Whilst Tse's approach could be considered quite original, it is difficult to see how his line of reasoning does any damage to Strawson's argument. All one has to do in order to reassert the consequences of the Basic Argument is to once again point to the omission from Tse's model of any examination of the process by which our brains "come to the decision" in the first place, before the criterial encoding takes place. In order for an agent to reach a decision based on sensory stimuli at T1, and thus set the appropriate criteria at T2, she must do so from within the parameters of her established psychology *as it is at T1* for it to be considered as *her* decision. And this, according to the Basic Argument, can only be a consequence of previous decisions or actions occurring at T0, T-1, T-2 and so on. All the mental work that the Basic Argument is focused on, therefore, occurs *before* Tse's criterial causation comes into play.

At one stage Tse appears to acknowledge these issues but then still tries to reaffirm that his theory adequately allows for decisions that are neither determined nor random by re-invoking indeterminism and the fact that the criteria may or may not be met by inherently variable inputs.

> If one pushes choices back far enough in the history of the system, those choices will have to be rooted in criteria that were not themselves selected by the system. We have to be born with a minimal set of criteria to get the whole process of criterial causation going. All later criteriality that is contingent upon a particular history and set of later decisions needs a starting point of criteria that were innately dictated by a genetic plan. This does not mean that there is no free will. It means that free will is constrained. We are not utterly free to choose the physical grounds of making a present choice. But our brains can make choices that are neither predetermined nor purely random, which are in part determined by previous decisions of our nervous system. The key point is that criteria will be met in unpredictable ways if there is inherent variability or noise in inputs, such as can be introduced by the randomness inherent in neurotransmitter molecules crossing the synapse or NMDA receptor behavior that can introduce chaotic behavior into the timing of neural spike trains. (Tse 2013, p.143)

Whether or not Tse is correct concerning the likely pattern of development from an initial "genetic plan,"[66] the addition of indeterminism does not appear to improve his position when it comes to meeting my (and his own) criteria for a libertarian free will. What it seems he is actually arguing for here is a model in which human action is a product of pre-determined parameters born from an established psychology, which is then rendered unpredictable by indeterministic inputs. Once again, he is trying to argue that adding deterministic and indeterministic processes together is the same as offering a "middle path." In fact, the result for the agent could arguably be described as them having *less* freedom from within this combination model than she would have from within a solely deterministic *or* a solely indeterministic model as she would lose access to factors supporting agential control in the former and factors supporting true ontological freedom in the latter. As mentioned already on a number of occasions in this section, Tse clearly states that he is arguing for a "strong free will" – one seemingly akin to my Gold Standard, involving both alternate possibilities (AP) *and* ultimate responsibility (UR) – but the above quote seems to indicate that what he is really arguing for, may actually be the more limited free action variety available to the compatibilists.

[66] I discuss this issue further in the conclusion (section 4.2).

Tse, it could therefore be argued, is not actually refuting the Basic Argument but instead trying to side step it by redefining what he means by free will, contradicting his own initial criteria in the process. The fact that, as he states, “All later criteriality that is contingent upon a particular history and set of later decisions needs a starting point of criteria that were innately dictated by a genetic plan,” is arguably the whole point of the Basic Argument: that every decision is made from a particular psychological standpoint established beyond our control, and thus our wills cannot be free in any sense relevant for responsibility (because responsibility, on this argument, requires we can meet the UR condition as well as the AP condition). On Tse’s own model, either the main decision is made deterministically from within this standpoint (from the agent’s CEP) prior to the encoding of criteria such as “stop and help,” in which case the decision to intervene or not is effectively determined by circumstances beyond the agent’s control, or the main decision is not made prior to the criterial encoding and what is encoded is as generic as “decide whether to intervene or not,” in which case the outcome is entirely random and dependent on the indeterministically variable inputs, and again beyond the agent’s control. Placing great emphasis (as Tse does) on the fact that in his model criteria are being set up in the *present* to influence *future* firing also does not really gain him anything. Positing a supervening (yet entirely physical) mental event capable of influencing future physical/mental/informational events and claiming this solves the problems the Basic Argument raises just does not follow. The actual decisions for what such mental events wish to make happen in the future remain subject to the ever-problematic *Catch-22 of Libertarianism.*

3.2.4 The *Criteria*

As discussed at the beginning of this section (3.2), Tse has set out the following conditions that he believes must be met in order for human beings to realise what he terms a “strong free will.” They are as follows:

1. Multiple courses of action must be genuinely open to the agent.
2. The agent must be able to choose among them.
3. The agent must be able to have chosen otherwise.
4. The choice must belong to the agent and not be solely random.

Tse believes his criterial causation, coupled with indeterminism and a non-epiphenomenal consciousness, can meet these conditions, presumably in the following ways:

1. Multiple courses of action must be genuinely open to the agent.
    a. The indeterministic nature of the universe means that a variety of different possible paths are genuinely available to the agent.
2. The agent must be able to choose among them.
    a. The agent has the ability to consciously assess information and select appropriate behaviours and the capability to rewire the neurons of her brain in order to encode the relevant criteria to achieve said behaviours.
3. The agent must be able to have chosen otherwise.
    a. The inherent variability of the synaptic inputs means just because one course of action was selected it does not mean another related course of action could not be realised were we to wind back the universe (thus satisfying the AP condition).
4. The choice must belong to the agent and not be solely random.
    a. It is the agent that carries out the internal deliberations and encoding the relevant criteria into our neuronal circuits (thus satisfying the UR condition).

In summary, Tse has argued that it is possible to have a libertarian free will that allows for non-determined but self-selected actions. Neurons are not simply blunt stimulus-response devices, solely at the mercy of their presynaptic inputs, but have the ability to recode their own response to stimuli (and that of other neurons around them) altering the informational criteria for when they will or will not fire. Patterns of neuronal activity harnessing this ability can therefore be causal in a way over and above the physical basis in which such patterns are ultimately realised and it is in this way that an agent can exert a form of top-down mental causation. Information, albeit realised in physical activity, can be causal of future information that will be realised in future physical activity. When we add indeterminism into the picture so that the inputs received are inherently variable and unpredictable, and the ability to consciously manipulate the contents of our endogenous attentional operations in deciding what criteria to encode in the first place, we have, according to Tse, all the ingredients we need for a strong (Gold Standard) free will.

I will now assess Tse's theory against my *Criteria for Coherence*. The first two principles can be taken together.

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.
P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.

Tse's theory is indeed (overall) an incompatibilist theory. Though he states the criterial causation part of his model would add biological weight to any compatibilist theory, he believes the addition of indeterminism is required to allow for a strong free will and thus argues extensively for indeterminism as the likely state of the universe, and incorporates it prominently into his theory. Thus, I would argue the first two principles have been met.

P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.

Initially, it might seem clear that Tse's theory also conforms to the AP condition as it states that, although criteria are preset in advance of the arrival of the inputs, the inputs themselves are unavoidably variable due to indeterminacies being influential at the synaptic level and thus an agent could choose a particular course of action but then, were the universe to be wound back, could chose a different course of action. Tse's example of thinking of a politician would seem to conform to this as a thought of Margaret Thatcher could just as easily be a though about Barack Obama. However, in other instances it seems more specificity is being encoded by the criteria, which puts complete conformation with this principle in doubt. As we discussed when applying Tse's model to our Jessica example, if the only undetermined variation available to her is *how* to intervene, not *whether* to intervene, it could be argued this is only a very limited form of available alternatives and not in keeping with the spirit of the AP condition. Thus, it is not clear that this principle has been met.

P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.

Tse would certainly argue that this criterion has been met, but it is not clear this is actually the case. Tse's model potentially sheds light on how an agent can exert some form of top-down control over their neural architecture in order to realise desires and attain goals through criterial encoding, but such encoding (as argued above) is presumably enacted *post-decision*. The decision first needs to be made and needs to be based in the agent's CEP in order for it not to be considered solely arbitrary. There is no discussion concerning exactly how the decision/desire is arrived at in the first place, leaving us again with such a state being either deterministically generated or arising randomly. Plus, the inclusion of indeterminism by way of random neuronal inputs arriving at preconfigured synapses brings into question precisely *how much* control Tse's model is providing the agent in any case. Ultimately, with Tse's model apparently featuring deterministic structuring followed by an undetermined arbitrating input, it could be argued that his model for decision-making is actually *part* determined and *part* arbitrary. Thus, I cannot say this principle has been met.

P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.

Tse's general answer to the problem of self-causation and regress is to rely on the notion of his criteria affecting *future* as opposed to present neuronal firing, in conjunction with the (at least partial) causal break posited by the prominent role of undetermined inputs. This, however, does not remove the question of the seemingly inevitable regress created in the actual decision-making or desire formation process. As we have seen in our discussion of Tse's model (and similarly with the other theories considered so far in this thesis), the problem just shifts to the relevant processes occurring *before* criterial causation takes place. As such principle five has not been met.

P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

Though in many ways rigorously scientific, there are a number of assumptions made and concepts included which would likely be considered contentious by some and Tse's apparent reliance on emergent phenomena, causally efficacious higher-level patterns of activity, and a supervening active consciousness could be considered unnecessary invocations of mystery. In addition, Tse's arguably ambiguous usage of terms such as "informational" and "mental" lend further obscurity to his model, resulting in his meeting of this principle being cast into doubt.

3.2.5 Conclusion

In conclusion, as judged against the *Criteria*, Tse's model for the Gold Standard libertarian free will cannot be considered coherent. Whilst an interesting neurological model for how our brains might operate, Tse's combination of criterial causation and indeterministic inputs produces a process of mental causation which, instead of being *neither* determined nor arbitrary, manages in reality to be a combination of *both*. In addition, despite claiming he is presenting a theory that provides for a "strong free will" Tse's theory really only presents a mechanistic procedure for the *implementation* of an agent's desires (more, free *action* than free *will*), with little or no discussion concerning key parts of her decision-making process and how such a process could avoid the *Catch-22 of Libertarianism*. Lastly, Tse's equivocations concerning the mental and physical aspects of human ontology and lack of clarity over key terms such as "information" or "informational" when discussing neural states adds an unnecessary level of obscurity to his model. Ultimately, Tse sets himself the task of meeting almost the same criteria by which I have judged his approach and has failed to meet both sets.

Most likely the key benefit of Tse's work is providing a useful neurological model for the implementation of decisions that philosophers such as Kane could draw on in support of their theories; though without offering any solutions to the key issues such libertarian positions face. Scientists (and neuroscientists in particular) undoubtedly have an important role to play in discussions in key areas of philosophy of mind – possibly also including those focussing on freedom and responsibility – but, as demonstrated here, they do not (thus far) appear to have helped provide any kind of concrete solution to the persisting problems. In the more scientific approaches, reductionism is often unavoidable and (on the face of the discussion in this chapter, at least) such a strategy is arguably unlikely to have much success in tackling the *Catch-22 of Libertarianism*. However, it could also be argued that the reductionist neuroscientific approach does not *go far enough* and, if all brain chemistry is ultimately realised in the basic building blocks of physics, perhaps what we need to do is seek an even *more* reductive analysis, involving not neurons, action potentials and NMDA receptors, but photons, electrons, and quarks. It is to such approaches that I shall now turn.

3.3 Roger Penrose and Stuart Hameroff's Orchestrated Objective Reduction

Anaesthesiologist Stuart Hameroff's paper "How quantum brain biology can rescue conscious free will" (2012) is based on his Orchestrated Objective Reduction (Orch OR) theory of consciousness, which he has developed with the mathematician Sir Roger Penrose. It is Hameroff's contention that, in their attempt to explain the origin and place of consciousness in the universe, he and Penrose have developed a theory that can not only offer a mechanism by which consciousness is generated (and show why it should not be considered epiphenomenal) but which can also account for the autonomous causal agency required for the Gold Standard freedom of the will. Consciousness, they argue, involves non-algorithmic processes which prevent its activity being solely deterministic and thus their model can provide the conscious agent with a mechanism by which they can have a direct controlling role over the biological workings of their own brain. As Hameroff states in the paper's abstract: "…Orch Or can account for real-time conscious causal agency, avoiding the need for consciousness to be seen as epiphenomenal illusion. Orch OR can rescue conscious free will." (Hameroff 2012, p.1)

In what follows, I shall explore and evaluate the Orch OR theory and consider whether it offers any new inroads into the problems faced by libertarian theories. I will then assess it against my *Criteria for Coherence* to judge whether it can be considered coherent.

3.3.1 OR, Orch OR and Microtubules

Objective Reduction

Objective Reduction (OR) is Penrose's own theory for the mechanism by which a quantum state that is evolving unitarily as a superposition of possible states in line with Schrödinger's wave equation, reduces to a single classical reality: often referred to as the "collapse of the wavefunction".[67] OR is in competition with other interpretations such as the Copenhagen Interpretation, Many Worlds theory or

[67] These processes are notoriously difficult to interpret (and even comprehend) but, in very simple terms, the quantum wavefunction is the evolution over time of the different probabilities associated with quantum states. These different states exist in what is known as a quantum superposition of simultaneous possibilities until a measurement is made, at which point the wavefunction "collapses" into a single reality.

Hidden Variables theory and states that reduction occurs once an objective physical threshold of gravitational (space-time) separation has been reached.[68]

> In the DP[69] version of OR, the reduction R of the quantum state does not arise as some kind of convenience or effective consequence of environmental decoherence, etc., as the conventional U formalism would seem to demand, but is instead taken to be one of the consequences of melding together the principles of Einstein's general relativity with those of the conventional unitary quantum formalism U, and this demands a departure from the strict rules of U. According to this OR viewpoint, any quantum measurement—whereby the quantum-superposed alternatives produced in accordance with the U formalism becomes reduced to a single actual occurrence—is a *real* objective physical process, and it is taken to result from the mass displacement between the alternatives being sufficient, in gravitational terms, for the superposition to become unstable. (Hameroff and Penrose 2013, p.51)

Penrose's OR is suitably detailed and complicated, but it contains two elements which are key for our purposes. Firstly, it involves at least some processes which Penrose argues are entirely non-computational or non-algorithmic (the terms computational and algorithmic are used interchangeably, essentially to mean anything that can be simulated on a computer).[70] Secondly, each reduction of the quantum state is accompanied by what is described as a discrete moment of "proto-consciousness" – an element of experience and awareness which, although in its most ubiquitous form is considered primitive and non-cognitive, has the potential to be harnessed within a wider scheme that ultimately results in full human consciousness and causal control of behaviour. Precisely *how* such moments can be coordinated (or rather, "orchestrated") to result in full consciousness I will consider next, but what is clear is that Penrose and Hameroff see consciousness as baked into the very fabric of the universe.

[68] This issue is the well-known *measurement problem* in quantum mechanics, summed up helpfully in Hameroff and Penrose 2013 as follows: "…the measurement problem is the conflict between the two fundamental procedures of quantum mechanics. One of these procedures, referred to as *unitary evolution*, denoted here by U, is the continuous deterministic evolution of the quantum state (i.e. of the wavefunction of the entire system) according to the fundamental *Schrödinger equation*, The other is the procedure that is adopted whenever a measurement of the system—or *observation*—is deemed to have taken place, where the quantum state is discontinuously and probabilistically replaced by another quantum state (referred to, technically, as an *eigenstate* of a mathematical operator that is taken to describe the measurement). This discontinuous jumping of the state is referred to as the *reduction* of the state (or the 'collapse of the wavefunction'), and will be denoted here by the letter R. This conflict between U and R is what is encapsulated by the term 'measurement problem'…" (Hameroff and Penrose 2013, p.50).

[69] DP here refers to Diosi-Penrose, reflecting the contribution of the physicist Lajos Diosi.

[70] This idea has arisen out of Penrose's arguments against the prospect of a strong artificial intelligence and its perceived algorithmic consequences for human conscious reasoning. For more on this see Penrose 1999 and 1994.

Orchestrated Objective Reduction

As discussed previously when considering the work of neuroscientist Peter Tse, the prevailing view of how our brains operate has tended to focus on the brute actions of the neurons, which then often leads to accusations that the entire process is solely algorithmic in nature and any emergent consciousness would likely be epiphenomenal and non-efficacious; which then further raises difficult questions concerning free will. Tse's attempt to address this issue was to focus on the capacity of neurons to reorganise themselves in order to implement desired behaviours and not just passively respond to inputs. Penrose and Hameroff, who also view the algorithmic conception of neural processes as a threat to free will, go somewhat deeper into the physical makeup of the neurons in order to look for gaps in the seemingly ballistic processes which could be exploited – though not in a way that would be considered solely random. As above, the suitable gap in current physics that Penrose and Hameroff have identified (and where they believe entirely new physical theories could be developed) is the seemingly discontinuous process of quantum-state reduction, to which they have added an associated and fundamental element of consciousness. What is needed in order to turn such instances of proto-consciousness into a fully functional working model is a mechanism by which the individual instances can be coordinated into any kind of coherent "whole," which could then act as some form of controlling entity. This is where the "orchestrated" part of Orch OR features.

Prior to any reduction the quantum state evolves unitarily (as per the Schrödinger equation), which involves the relevant quantum computations associated with the various superposed states. This process then terminates by Penrose's OR once the required gravitational threshold is met. It is this process of quantum computation, prior to reduction, that Penrose and Hameroff argue needs to be "orchestrated" in order for any form of agent control to enter the system but, for this to occur in the brain, the relevant superposed state needs to be maintained for long enough without it spontaneously collapsing due to environmental decoherence (standard ORs would collapse randomly and rapidly preventing any real orchestration). Again as mentioned when discussing Tse's model, the prospect of quantum effects being relevant in such large structures as neurons and synapses is a contentious issue, with many arguing that such warm and wet conditions as found in a human brain would be a very hostile environment for maintaining fragile quantum states. As such, Hameroff's main contribution to the theory has been to identify a biological location which could provide a more friendly environment in which quantum computations could progress successfully: the microtubules.

Microtubules form part of the cytoskeleton of neurons and other cells and perform various different functions. They are cylindrical polymers of tubulin, only 25 nanometres in diameter, and with lengths believed to vary from a few hundred nanometres to perhaps up to metres in some nerve cells. Importantly, microtubules appear to be particularly prevalent in neurons and remain stable because neurons do not divide like other cells. They are also, according to Hameroff, believed to be arranged in such a fashion as to create networks suitable for information processing (with the related potential to vastly increase brain capacity) as well as stable enough for encoding memories and information over the longer term. Hameroff discusses how neuronal microtubules play a role in influencing axonal firing and regulating synapses and synaptic plasticity, which is what would appear to be the bio-physical link between the superposed quantum computation and the biological activity that ultimately controls behaviour: "Each OR reduction selects microtubule states which can trigger axonal firings, and control behaviour" (Hameroff 2012, p.1). Furthermore, by utilising available gap junctions much larger entangled networks can be achieved across the brain, creating and maintaining a much greater number of microtubules in a state of superposition; and there is evidence that – due to the structure and arrangement of the microtubules – such a quantum-superposition could indeed be maintained for the required length of time, even at a normal brain temperature.[71]

As with the physics, the biology is detailed and complex, but the important point for our purposes is that there would appear to be evidence that microtubules could provide a suitable biological environment in the brain, capable of both maintaining quantum superposed states for long enough to allow for any required orchestration (influenced by synaptic inputs, memory and other cognitive functions) that brings in non-computable (but also non-random) elements and is also capable of directing/manipulating synaptic activity once the OR threshold has been met and the state collapses. OR without the orchestration (without the *Orch*) would just be at the mercy of its environment and subject to random elements, but when orchestrated ("adequately organized, imbued with cognitive information, and capable of integration…" Hameroff and Penrose 2013, p54) and kept isolated from random environmental influences for long enough Orch OR can occur fully, which results in moments of consciousness and allows for agent control.

[71] Hameroff cites research carried out by Anirban Bandyopdhyay and colleagues, who, he states, do seem to have provided "clear evidence for coherent microtubule quantum states at brain temperature." (Hameroff and Penrose 2013 p.55)

> *Consciousness results from discrete physical events; such events have always existed in the universe as non-cognitive, proto-conscious events, these acting as part of precise physical laws not yet fully understood*. Biology evolved a mechanism to orchestrate such events and to couple them to neuronal activity, resulting in meaningful, cognitive, conscious moments and thence also to causal control of behavior. These events are proposed specifically to be moments of quantum state reduction (intrinsic quantum "self-measurement"). Such events need not necessarily be taken as part of current theories of the laws of the universe, but should ultimately be scientifically describable…In the Orch OR theory, these conscious events are terminations of quantum computations in brain microtubules reducing by Diósi–Penrose 'objective reduction' ('OR'), and having experiential qualities. In this view consciousness is an intrinsic feature of the action of the universe. (Hameroff and Penrose 2013, p.39)

3.3.2 Backwards Time Referral

Hameroff and Penrose specifically discuss the famous experiments carried out by Benjamin Libet and his followers – which many believe are a real stumbling block for free will – in order to demonstrate the potential benefits of the application of quantum theory to human actions. Libet's experiments, which focussed on recording electrical activity in the brain of their subjects whilst also asking them questions concerning their conscious intentions, seem to show that conscious perceptions occur, in fact, only *after* a decision to act has been taken physiologically, ruling out any conscious causal efficacy and casting doubt on the possibility of genuinely willed action (except perhaps the capability to consciously reverse or overrule a non-conscious impulse to act).

Penrose and Hameroff argue contrary to this that quantum theory shows it is in fact possible to refer quantum information *backwards* in classical time (in their case due to the temporal non-locality caused by OR), which could account for the Libet's findings and potentially restore an efficacious consciousness. Subjects in the Libet experiments do not falsely remember acting consciously but, rather: "…a quantum explanation with temporal non-locality and backward time referral enables constructive "filling in" from near future brain activity, allowing real time conscious perception" (Hameroff 2012, p.7). It is in this way, according to Hameroff, that our consciousness remains the true arbiter of our willed actions, retaining its capacity to regulate axonal firings and thus coordinate real-time behaviour even if it is in some sense sending quantum information backwards in classical time to do so.[72]

Penrose discusses such temporal anomalies extensively in his various works, noting that the experience of the passage of time – the notion of time as something that 'flows' from the past to the future – does not always align with what our best physical theories have to say on the matter. There is nothing about time *flowing* in either relativity theory or quantum mechanics, which raises the question of how meaningful it is to talk of a conscious event and associated physiological activity being seemingly required to occur in any particular temporal order. Penrose speculates that when dealing with non-local quantum entanglements it may be that quantum information can actually propagate in either direction in time because these "quanglements" (as he calls them) are not mediated in a normal causal way. Quantum information going backwards in time (or forwards in time, for that matter) is taken to be acausal as it cannot send signals – as this would lead to all sorts of problematic causal paradoxes – but it does not rule out any form of influence whatsoever.

> The issue is a subtle one, but if conscious experience is indeed rooted in the OR process, where we take OR to relate the *classical* to the *quantum* world, then apparent anomalies in the sequential aspects of consciousness are perhaps to be expected. The Orch OR scheme allows conscious experience to be *temporally non-local* to a degree, where this temporal non-locality would spread to the kind of timescale that would be involved in the relevant Orch OR process, which might indeed allow this temporal non-locality to spread to a time of Libet's 500 milliseconds ('ms') or longer. When the 'moment' of an internal conscious experience is timed externally, it may well be found that this external timing does not precisely accord with a time progression that would seem to apply to internal conscious experience, owing to this temporal non-locality intrinsic to Orch OR. The effective quantum backward-time referral inherent in the temporal non-locality resulting from the quanglement aspects of Orch OR, as suggested above, enables conscious experience actually to be *temporally-nonlocal*, with backward time effects seen as temporal variability in axonal firing threshold…consciously regulating behavior and providing a possible means to rescue consciousness from its unfortunate characterization as epiphenomenal illusion. Accordingly, Orch OR could well enable consciousness to have a causal efficacy, despite its apparently anomalous relation to a

[72] Hameroff in fact states that Libet himself argued for the possibility of subjective information being referred backward in time in order to save an efficacious consciousness from his own experiments. He states that Libet's assertions were "disbelieved and ridiculed…but never refuted." (Hameroff 2012, p.7)

> timing assigned to it in relation to an external clock, thereby allowing conscious action to provide a semblance of free will. (Hameroff and Penrose 2013, p.64)

3.3.3 Jessica and her Time Travelling Consciousness

To summarise (and overly simplify) the Penrose-Hameroff Orch OR theory: quantum activity within microtubules (and via gap junctions) creates a vast network of entangled particles in superposed quantum states, which evolve in line with the Schrödinger equation until the threshold for Penrose's (non-computational, gravitational) OR is reached, at which point the wavefunction collapses, a new and unknown process takes place and a moment of "proto-consciousness" occurs. When such instances are suitably *orchestrated*, these combine to produce the sort of subjective consciousness associated with phenomenal experience and intentionality (integrated with memory) which can then influence neuronal firings/inhibitings (potentially) controlling bodily actions. The backward referral in time of quantum information is proposed to prevent consciousness being rendered epiphenomenal and to prevent this being either a completely algorithmic or a completely random process, thus providing the created consciousness with an efficacious rather than solely an observing role and thereby apparently rescuing free will from both algorithmic inevitability and random environmental influences.

Penrose and Hameroff never fully extrapolate their model up to the level of visible human action but, were we to apply it to our athlete Jessica's dilemma, it might proceed as follows: upon seeing the elderly man fall to the ground, perceptual information would filter through Jessica's brain, initiating quantum computational activity within neuronal microtubules which would spread out across relevant regions of her brain through entanglement via gap junctions. The quantum state would be maintained within the isolated environment for a sufficient amount of time for a level of orchestration (Orch) to occur; presumably involving her perceptions, memory, and a variety of other cognitive applications generated via her current established psychology (CEP) which can in some way influence the computation process. At some point after the superposition has been maintained long enough for some kind of agential involvement/guidance, a threshold of sufficient gravitational separation between the superposed quantum states is reached and the quantum computations terminate by Penrose's objective reduction (OR) – a new, unknown, non-algorithmic (but ultimately physical), process accompanied by conscious awareness. The OR reduction (in a sense) *chooses* the classical microtubule state Jessica's brain "collapses" into, which becomes the output of the quantum computation and which then triggers axonal firings which ultimately cause Jessica to, say, exit her vehicle and assist.

Some apparent issues arise when attempting to incorporate the Orch OR theory into a psychological-level example of actual decision-making and action which do not appear to be resolved in the current literature on the theory. Firstly, it is not completely clear how the "discrete conscious moment" is related to the objective reduction. At times it is stated that the OR is "accompanied" by a conscious moment, at other times the words used are "associated with," and at other times it appears the conscious moment is *identified with* the OR.[73] This makes it difficult to utilise the concept in any kind of "real-world" example of cognition. In fairness, Penrose and Hameroff do appear to acknowledge the unknowns present in their theory:

> It is to be expected that the actual mechanisms underlying the production of consciousness in a human brain will be very much more sophisticated than any that we can put forward at the present time, and would be likely to differ in many important respects from any that we would be in a position to anticipate in our current proposals. Nevertheless, we do feel that the suggestions that we are putting forward here represent a serious attempt to grapple with the fundamental issues raised by the consciousness phenomenon, and it is in this spirit that we present them here. (Hameroff and Penrose 2013, p.58)

Secondly, there is not a great deal of discussion concerning what is actually occurring during the orchestrated quantum computational process. It is not clear how such a process might be influenced in any way by the agent's CEP, in order to allow for sufficient control and meet the ultimate responsibility (UR) condition. Thirdly, at times the orchestration process itself appears to relate to the quantum computations occurring prior to OR, at other times it appears to relate to the requirement to orchestrate the conscious *results* of the OR, and at other times it seems to relate to a combination of both.[74] Given the vast numbers of neurons involved in even the most basic kind of cognition (not to mention the immense complexity of the quantum phenomena involved), I suspect the orchestration concept would indeed have to apply both to the quantum computation phase and the resulting moments of consciousness, but the lack of clarification on this point is one of the reasons it is a difficult

[73] See Hameroff and Penrose 2013, pages 53, 59, and 67 for examples.
[74] See Hameroff and Penrose 2013, pages 59, 71 and 73 for examples.

theory to apply at any kind of psychological level. Lastly, whilst clear that the OR process contains non-computational elements (and thus would perhaps go some way to meeting the AP condition), it is not similarly clear if the prior quantum state evolution, even when orchestrated, is entirely algorithmic or not. At times the suggestion appears to be that it is (in line with the Schrödinger equation) at other times it appears the orchestration/cognitive influence, might include some non-algorithmic elements.

Concerning the voluminous scientific detail presented in the various papers, a number of specific objections have been raised against various aspects of the Orch OR theory[75] but the important questions for my purposes (which have not yet been addressed) are as follows. Even if we were to try and ignore the immediate issues just raised above and were to accept that as a scientific theory for the mechanism of consciousness generation and neurological causation the Orch OR theory is correct, can it make any progress against the key objections I have been considering – the Luck Objection and the Basic Argument? Does the theory offer any potential escape from the identified *Catch-22 of Libertarianism*? Can it meet my proposed *Criteria for Coherence*? It is to these questions I shall now turn.

3.3.4 The Luck Objection and the Basic Argument

The Luck Objection

Penrose and Hameroff argue that the important addition of a layer of "orchestration" to their model (allowed for by the maintenance of the quantum-superposed state for a sufficient length of time) prevents the process from being overwhelmed by random factors, as they state it *would* be were the state to decohere quickly due to entanglement with its environment.

> The role of environmental decoherence, according to OR schemes, is that R is effected in a system through its entanglement with its much larger effectively random environment, so that when OR takes place in that environment, the system itself is carried with it and therefore reduces, seemingly randomly, in accordance with a conventional R process. Thus, if we require non-random aspects of OR to play a role in (conscious) brain function, as is required for Orch OR, we need to avoid premature entanglement with the random environment, as this would result in state reduction without non-computable aspects or cognition. For Orch OR, environmental interactions must be

[75] See their own roundup of the objections in Hameroff and Penrose 2013, p.66.

> avoided during the evolution toward [reduction threshold], so that the non-random (non-computable) aspects of OR can be playing their roles… (Hameroff and Penrose 2013, p.62)

Indeed, they state that the vast majority of OR events are *not* orchestrated as part of some coherent structure so would be overwhelmed by random factors and, as such, would have no significant conscious experience associated with them. These would only produce their primitive "proto-conscious" events (without information or meaning) which, on their own, only form the isolated ingredients for a fully functional and experiential consciousness. When they *are* orchestrated, however, the randomness disappears. What is not clear, though, is just *how* this is achieved. It is argued that quantum computations progress within the microtubules, after which OR (which likely requires some non-computational new physics to fully explain it) acts to "select" specific states. But what precisely is guiding such a selection, is not discussed in any detail.

As discussed in the previous sections focussing on other theories, the main thrust of the Luck Objection is that if we believe that having genuine alternative possibilities (in an indeterministic sense – thus meeting the AP condition) and having the ability to choose between them (in an autonomous fashion undetermined by past events) are essential requirements for free will, we are lead to the conclusion that in order to satisfy these requirements we have to allow for an agent, at any point in time, to be able to both act or act otherwise, even with all past circumstances being exactly the same. With nothing in an agent's past or present apparently capable of determining their actions, the argument is that what course of action *is* ultimately decided upon can only be down to luck.

One problem with assessing the Orch OR theory against the Luck Objection is that it is not clear from the presentation of the theory exactly what parts of the process may or may not be considered to be deterministic or indeterministic. At some points in the literature it is suggested that the quantum computation process that occurs prior to OR proceeds entirely deterministically in line with the Schrödinger equation. At other points it is suggested that cognitive influences during this computation could add an important "x-factor," suggesting the process may not be entirely determined. If there is to be any indeterministic influence, it would seem that the most likely stage for its involvement would be during the OR process but, if this is the case, then this raises familiar issues concerning how such a process could come under the agent's control. What is clear, is that Penrose and Hameroff do see their model as providing a mechanism which can indeed navigate the difficult path between determinism and randomness. As Hameroff states: “Thus conscious choices in OR (and Orch OR) are neither random nor algorithmically deterministic” (Hameroff 2012, p.11).

If we refer back to the Jessica example, we can seek to identify key points in the process at which accusations of luck and randomness are most likely to be levelled. After perceptual input and presumably some form of initial cognitive assessment we likely enter the key quantum computational phase. During this phase, in order for the procedure to count as a fully orchestrated process which would satisfy the conditions to result in a moment of consciousness, other cognitive influences (Jessica’s CEP) are apparently also brought to bear. The extent of the involvement of the CEP in this phase is unclear (as is whether it is during this phase that the decision would be *made*, or whether that would occur during the OR) but if the CEP’s involvement is essentially *determining* then Orch OR does not seem to escape from the algorithmic/deterministic issues its authors have identified as a threat to free will (relating both to the influence of the CEP over the decision making process and the formation of the CEP itself). However, if the involvement of the CEP is *not* determining then we are back to familiar issues (particularly when considering the work of Robert Kane) of what then would “tip the balance” in favour of one choice over the other.

There is then the matter of the subsequent OR, which is stated not to be well understood but believed to be a non-computational process that results in a discrete moment of consciousness. Given the claims that the OR process is not algorithmic in nature (and thus not “computable” from the past), this stage would perhaps seem more vulnerable to charges of luck and randomness. Presumably, if the OR is not to be considered completely disconnected from the preceding (CEP influenced) quantum computations – and thus entirely random – then it must in some way be influenced by them, but then we risk falling back into algorithmic territory. And what of the associated conscious moment? Penrose and Hameroff are clear in their regular use of terms such as “conscious causal control” and “conscious free will” that consciousness plays a key role in free will, but its location in the process – coming after the CEP influenced quantum computation – and the potentially disjointed way in which it is *created* by the OR raises doubts over its causal efficacy. The inclusion of the concept of backwards time referral, potentially utilised by consciousness to send quantum information back to influence the quantum computations occurring prior to OR, does not appear to solve the problems of agent-control as there would remain a point (or an interval) at which conscious decisions would have to be made prior to the information being referred back and all the issues raised above concerning determinism or randomness could just be applied to that stage in the process.

If the moment of consciousness is *identified with* OR, rather than just accompanying it in some disjointed sense (as the literature at some points seems to suggest),[76] then perhaps it could be said that the (CEP influenced) quantum computation terminates in a conscious moment which acts as the moment of "decision," the OR counterpart of which then selects the appropriate microtubule states which initiate the relevant axonal firings, and so forth. The question then becomes whether the resulting conscious decision is "informed" or "necessitated" by the preceding quantum computation. If it is not necessitated by it, and thus the decision could equally have been different, then the Luck Objection appears to become valid once again. The quantum computations carried out across networks of entangled microtubules in Jessica's brain (informed or orchestrated by her CEP) could incline fully towards ignoring the fallen male and driving away, which could then be precisely what is initiated by the OR/consciousness step. However, the argument remains that if we were to wind the universe back to the instant just prior to the OR and let it play forward again, if the computation process cannot be determining then the computations that inclined her towards driving away could still result in an OR/consciousness phase which instead causes her to ultimately get out and help.

Rather than answer the Luck Objection, the Orch OR model (at least as extrapolated up to the level of psychology here) seems likely to contain a good deal of random elements and disconnection between key phases of the model and any attempts to remove either leads back to the algorithmic/deterministic consequences Penrose and Hameroff wish to avoid.

The Basic Argument

The crux of the Basic Argument, it will be recalled, is that in order for an agent's decisions to count as *hers* they must stem from her current established psychology (her CEP) in a presumably determined fashion but, in doing so, a vicious regress is created of a decision which is caused by a CEP, which is itself formed from previous decisions, which are themselves the product of a previous EP, and so on back into the past. The question, then, is whether the Orch OR model can break the regress in such a way that the

[76] See Hameroff and Penrose 2013, p.67.

agent could retain a satisfactory control over her actions and thus meet the UR condition. As in the previous section, much of the debate rests on exactly where in the quantum/cognitive processes described any causal breaks might come and also at what stage a decision is likely to be made, neither of which are clearly identified.

If the decision is essentially made during the quantum computation process – as orchestrated by the CEP – then the issue of regress arises in connection with both the unitary evolution of the quantum state *and* the formation over time of the CEP. If the decision is made by the conscious moment, then this is either fully informed by the previous computation or finalised at that moment in conjunction with the agent's CEP (the development of which can be traced back through the life-history of the agent); both scenarios again leading to a regress. If the conscious moment makes the decision and refers information backwards in time, the regress remains due to the required involvement from the CEP prior to the information being sent back. Granted the regress issue is slightly confused due to the convoluted timeline, but the principle remains.

As with the Luck Objection, the Orch OR model does not appear to be able to offer any new avenues of argument against (or response to) the Basic Argument.

3.3.5 The *Criteria*

I will now review the Orch OR theory against my *Criteria for Coherence*, starting (as previously) with the first two principles taken together.

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.
P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.

Sir Roger Penrose himself, it has to be said, at times appears noncommittal about both the existence of free will and whether or not the universe will turn out to be deterministic or indeterministic (his interest is more in arguing for a process which is non-computable – non-algorithmic – which does not *necessarily* mean non-deterministic),[77] but a close reading of his key texts shows a real focus on the importance of human creativity and inspiration as well his placing great emphasis on the suggested non-computational elements of consciousness (via the OR process), which would perhaps indicate an inclination toward an incompatibilist view on human freedom. Hameroff is much more explicit in this area, using Penrose's arguments against a purely computational consciousness as a basis to dismiss physical determinism and epiphenomenalism (the existence of both – or either – of which he believes would preclude free will) and claiming that the Orch OR theory can account for conscious free will that is clearly of a libertarian variety. In addition, Penrose's own theory relies on an as-yet unknown, but definitely non-computational, quantum-mechanical process during the reduction of the quantum state, which would seem at least to make the case for a significant presence of indeterminism in the system – as many interpretations of quantum theory do more generally. Hameroff, once again, is more explicit and it is clear that he has designed the Orch OR model, which incorporates Penrose's non-algorithmic process, specifically to show that it is not just a deterministic combination of past circumstances and physical laws which account for all our actions. I would argue, therefore that both principle one and principle two have been met.

P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.

Again, Penrose and Hameroff would likely point to the fact that their model rests on a quantum-mechanical foundation (which inherently contains indeterministic/probabilistic elements) and also point to the non-computational elements of the proposed quantum-mechanical consciousness as evidence of agential freedom from any form of determinism; all of which means the model arguably meets the AP condition. As before, Penrose's position is unclear, but I believe that his rejection of the universal picture provided by classical physics would likely lead him to the conclusion that it was at least possible that a human agent could be free from any physical causal constraints. Principle three is therefore also met.

[77] "…this new procedure would contain an essentially non-algorithmic element. This would imply that the future would not be computable from the present, even though it might be determined by it. I have tried to be clear in distinguishing the issue of computability from that of determinism, in my discussions in Chapter 5. It seems to me to be quite plausible that CQG might be a deterministic but non-computable theory." (Penrose 1999, Chapter 10)

P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.

Though they never get into the specifics of decision making or elevate their model to the level of human psychology, Hameroff has stated clearly that he believes the Orch OR theory does provide a model for conscious choices which is neither random nor determined and fully under the control of the agent. It is argued that the model provides for an efficacious consciousness which can exert direct physical influence within the brain in a manner which is not solely the result of algorithmic processes and neither subject to entirely random environmental factors. However (as we have seen above), the model seems in fact to contain some processes which appear deterministic and some which appear essentially random. When considered against the Luck Objection the model could not account for how an agent could have full control over their decision to implement a particular course of action while resting on the indeterministic foundations of quantum mechanics. Principle four has therefore not been met.

P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.

As with potential elements of randomness, various potential regresses also appear prevalent in the Orch OR model. Wherever the main locus of agent-causal control is placed – be it during the CEP orchestrated quantum computations or the OR/Consciousness event – if the act is not to be entirely disconnected from its antecedent motivations, some form of problematic regress appears inevitable. The addition of the notion of consciousness referring quantum information backwards in classical time to in some way influence events does not help matters as it just shifts the faculty subject to the regress to a different part of the process. Principle five, therefore, is also not met.

P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

While Penrose and Hameroff would likely wish to distance themselves from the word “mystery” when presenting their theory, there are certainly (by their own admission) plenty of “unknowns.” This is perhaps understandable given that the subject matter focuses on providing a quantum-mechanical/neurological foundation for consciousness (the amalgamation of two subjects notoriously mysterious in many ways) but it could be argued that more work needs to be done in certain areas before the theory can really be put to test, philosophically speaking. The OR process alone likely requires a whole new physical theory to fully explain it and there is currently little or no discussion given to the ontological nature of their conscious moments (or proto-conscious moments) or to how such phenomena are created, or any detail provided on the important method by which cognitive orchestration takes place. However, Penrose and Hameroff would no doubt argue that the mysteries present are at least *scientific mysteries*, arising from an area of study largely in its infancy. They may point out that all scientific discoveries were once scientific mysteries and it is only through invention and speculation that science can ever move forward. The mysteries in their theory, they would say, are not *necessarily* mysterious but rather simply scientific facts waiting to be discovered. That being said, I still do not believe principle six has been met.

3.3.6 Conclusion

The Orch OR theory presents an intriguing model of how quantum mechanical operations in the brain could provide a foundation for cognitive processing and conscious control. At its core is no doubt solid scientific reasoning concerning Penrose's Objective Reduction as a competitor to other interpretations of the measurement problem in quantum theory and concerning Hameroff's belief that microtubules would provide a favourable environment for quantum computation in the brain. The problems come when these central notions are used to construct a framework for not only consciousness but also free action. It is at this stage the model becomes too vague in many of its assertions and this becomes particularly problematic when attempts are made to extrapolate the framework to the psychological level in order for it to be considered as a basis of a libertarian theory of free will.

The theory raises a host of questions as many of its key elements are unclear, including exactly what kind of entity consciousness is in this model, how or when the orchestration process occurs, what kind of psychological elements are likely to be involved, how they would be involved, whether the orchestration process is a deterministic process, whether the OR is a discontinuous event, how the orchestration process influences OR, how or why such an event results in moments of consciousness, and so forth. But the most crucial thing to note for our purposes is that even if we were to decide on answers for such questions, it seems apparent the theory would suffer from all the same problems that have beset the other libertarian theories considered up to this point. In the face of the Luck Objection, the theory cannot account for how either the preceding quantum computations or the OR/Consciousness event (wherever the locus of agent control might lie) could exert a satisfactorily non-determining control over an agent's actions. In the face of the Basic Argument, the model could not account for how an agent could exert a non-random control anywhere within the quantum mechanical framework in such a way as not to lead to an infinite regress. Lastly, the theory failed to adequately meet the *Criteria* and so cannot be considered coherent.

Penrose and Hameroff, however, are not the only scientists to apply their fundamental physical theories to the topics of consciousness and free will. A number of others have done likewise, and it is to one of these I shall tun next to see if their version of a quantum-mechanically inspired model for human thought and action can make better progress than the Orch OR model in producing a coherent libertarian theory.

3.4 Henry Stapp's Quantum Interactive Dualism[78]

Physicist Henry Stapp (similarly to Penrose and Hameroff) claims that by fully accepting quantum mechanics as the correct theory of our physical universe, we can develop a model for free will which conforms to the main tenets of the libertarian position. Stapp believes that neuroscientists and philosophers focusing on such topics as freedom and responsibility (as well as consciousness and other topics in Philosophy of Mind) often mistakenly base their theories on a foundation of classical (Newtonian) physics, which is only really an approximation to the *true* (quantum) physics. This practice then leads them to the inevitable deterministic conclusion that every event has causally sufficient antecedent conditions and, as such, leaves no room for free will in our world. "Orthodox" quantum mechanics based on John von Neumann's formulation of the Copenhagen Interpretation, however, (according to Stapp) not only allows for but *requires* an intervention by a conscious agent into the basic physical dynamics in order to resolve an evolving quantum superposition of possible states. Furthermore, such interventions are themselves not in any way fixed in advance – or even mathematically described – by the known quantum laws. The required intervention is said to confer sufficient control onto the agent and the absence of any calculable details concerning the nature of this intervention is argued to mean that the achieved control is in no way determined by prior conditions. Stapp would thus consider his model as potentially offering a way out of the ever-problematic *Catch-22 of Libertarianism*.

After a (very) brief discussion focussing on the switch from classical to quantum physics and Stapp's preferred interpretation of the quantum formalism, I will present and evaluate Stapp's model and assess it against my *Criteria for Coherence*.

3.4.1 From the Classical to the Quantum

Stapp is consistently clear in his view that a universe in which classical physics holds at the most fundamental level has no room in it for free will or any kind of efficacious mental events.[79]

[78] This is the title of Stapp's 2005 paper – and I believe a fitting name for his theory overall – but he does not seem to have retained it himself in subsequent publications.

[79] Stapp refers regularly in his work to both the mental and physical aspects of reality, without really being clear on what he takes the mental aspects to be, ontologically speaking. I discuss this further in the section's conclusion.

> The core precept of classical mechanics is this: The causal dynamical evolution of the physically described properties of nature is completely determined by physically described properties alone, with no reference to mental realities. This conception of nature reduces human beings to essentially mindless automatons: mental processes may indeed be happening, but they cannot influence in any way the evolution of the physically described universe! Mind is thereby made purely 'epiphenomenal'! (Stapp 2011, p.154)

Such statements are obviously highly contentious and raise a number of hotly debated questions concerning the ontological nature of mental events, their place in the physical world, and their connection (or lack of such) to any potential neural substrate. I do not intend to specifically address such issues here, however, because – as will become evident – even if we are to accept Stapp's view of such things as entirely correct, this does not affect *my* chosen line of attack against his consequent model for a libertarian free will.

Stapp's solution to this supposed rendering of our mental events as causally impotent by the precepts of classical physics is simply to argue that this theory has now been shown to be incorrect and has been replaced by a new theory, quantum mechanics, which has been empirically verified to great accuracy. Further, he argues that not only does this mean we have no need to base our models for free will on a Newtonian foundation of entirely determined physical interactions, but that we have a new quantum theory which actually *requires* an intervention by a conscious agent in order to reduce the continuum of quantum-superposed states into a single, classically describable reality. Further still, he argues that such an intervention is itself not in any way determined by the known quantum laws.

> The new theory departs from the old one in many important ways, but none is more significant in the realm of human affairs than the role it assigns to your conscious choices. These choices are not fixed by the laws of the new physics, yet these choices are asserted by those laws to have important causal effects in the physical world. Thus contemporary physical theory annuls the claim of mechanical determinism. In a profound reversal of the classical physical principles, its laws make your conscious choices causally effective in the physical world, while failing to determine, even statistically, what those choices will be. (Stapp 2011, p.VII)

This view is derived (specifically) from a particular formulation of Stapp's preferred interpretation of quantum theory: John Von Neumann's mathematical formulation of the Copenhagen interpretation.[80] The important point for our purposes is that Stapp considers this version of quantum mechanics to contain a significant "causal gap" in its basic dynamics which the quantum laws alone are not able to resolve, meaning the evolving spectrum of allowed superposed quantum states would just continue indefinitely. What is needed is an intervention by a conscious agent in (at the very least) the setting of new boundary conditions in such a way as to enforce state reductions that then begin a new process of evolution. In contrast to the classical picture of the universe – in which particles like tiny marbles travel along set paths through classical fields, determined from the time of the big bang and entirely independent of whether they are being observed by any human agents (which, on Stapp's view, would result in the exclusion of any kind of efficacious consciousness) – the quantum laws (though much more accurate in predictive practice) are fundamentally incomplete in the way just described. This leaves the need for some "extra factor" to intervene and reduce the quantum state to a single actuality which, Stapp argues, forced the founders of quantum theory to introduce the notion of the "free choice of the experimenter" (later termed Process 1 interventions by von Neumann). Without such interventions all the quantum laws can specify is a continuously evolving smear of probabilities of possible outcomes.

Key to Stapp's view, therefore, is his conception of quantum theory as ultimately *subjective* in that it is not so much about the actual behaviour of the quantum phenomena themselves but more about our *knowledge* of this behaviour and our role in interrogating the world around us. It is about human agents making predictions about observable outcomes, choosing experiments to perform, and then communicating their results. From this Stapp proclaims the theory to be "intrinsically psychophysical" (Stapp 2011, p.2) with both a physical aspect that refers to the mathematical representation of the physical world, and a psychological aspect that refers to our streams of conscious experiences and mental intentions and so forth. He claims there is a necessary causal link between the two that classical physics eliminated but which quantum theory restores.

[80] The Copenhagen interpretation of quantum mechanics is attributed to a group of physicists working in the early days of quantum theory, including Niels Bohr, Werner Heisenberg and Max Born. Though more a collection of views rather than a definitive interpretation, key features common to the views include the intrinsically indeterministic nature of quantum mechanics and, importantly for our purposes, the view that the only "truth" that can be attributed to a physical object or state is the "result" of some measurement or other – seemingly thus requiring an observer of some kind.

> …according to the new conception, the *physically described world* is built not out of bits of matter, as matter was understood in the nineteenth century, but out of objective *tendencies* – potentialities – for certain discrete, whole *actual events* to occur. Each such event has both a psychologically described aspect, which is essentially an increment in knowledge, and also a physically described aspect, which is an action that *abruptly changes* the mathematically described set of potentialities to one that is concordant with the increase in knowledge…The most radical change wrought by this switch to quantum mechanics is the injection directly into the dynamics *of certain choices made by human beings about how they will act*. Human actions enter, of course, also in classical physics. But the two cases are fundamentally different. In the classical case the way a person acts is fully determined in principle by the physically described aspects of reality alone. But in the quantum case there is *an essential gap in physical causation*. This gap is generated by Heisenberg's uncertainty principle, which opens up, at the level of human actions, a range of alternative possible behaviours between which the physically described aspects of theory are in principle unable to choose or decide. But this loss-in-principle of causal definiteness, associated with a loss of knowable-in-principle physically describable information, opens the way, logically, to an input into the dynamics of another kind of possible causes, which are eminently knowable, both in principle and in practice, namely our conscious choices about how we will act. These interventions in the dynamics take the form of specifications of *new boundary conditions*. (Stapp 2011, p.9)

As Stapp presents quantum theory, then, the evolution of the physically described quantum state (governed by the Schrödinger equation) can offer only a continuum (or "smear") of possibilities and cannot resolve itself into a single physical state that matches what is being experienced without some form of intervention. Given the "psychophysical" nature of quantum theory, this intervention is presumed by Stapp to be one by a conscious agent; akin to the choosing of an experimental arrangement that partitions which of a set of complementary types of phenomena one wishes to study. Such interventions are argued by Stapp to be "free" because they are not determined in any way by the physical part of the theory – not accounted for whatsoever in the preceding evolution of the quantum state. Such interventions bring conscious intentions into the physical dynamics but, whilst the *effects* of such choices may be fixed by quantum laws, the content or timing of the interventions themselves are not.

As above, there are no doubt numerous parts of Stapp's interpretation of quantum theory that people would wish to take issue with but, as my focus is on the search for a coherent theory of libertarian freedom, I am most interested in how Stapp uses this interpretation to formulate a model for free action; to which I shall now turn.

3.4.2 Processes 0, 1 2 & 3, Templates for Action, and the Quantum Zeno Effect

Process 0-3

Stapp contends that the so termed “free choice of the experimenter” (or “observer”) is not limited to actual scientists in a lab setting up equipment and probing quantum phenomena but is also intended more metaphorically and is applicable to all wilful efforts made by conscious agents from infancy upwards. By instigating their own conscious interventions (probings) agents can partition quantum states in such a way as to influence outcomes and achieve desired goals.

Stapp utilises von Neumann’s terminology when describing his own model for free will. “Process 2” is the term given to the evolution of the quantum state *between* interventions, as described by the Schrödinger equation. Such an evolution is entirely deterministic and mechanical and leads to the generation of a multitude of physically described possibilities. What Process 2 is unable to do is resolve this ever-growing superposition of possible states by reducing it to the type of singular classical reality that corresponds with our actual experiences. The required intervention to help achieve this state-reduction is named “Process 1” and is described by Stapp as: “the basic probing action that partitions a potential continuum of physically describable possibilities into a (countable) set of empirically recognizable alternative possibilities” (Stapp 2011, p.24). It is then from this set that one is selected to potentially become actualised. Adding to von Neumann’s model, Stapp identifies two further important processes, which he calls “Process 3” and “Process 0” (zero). Process 3 – which he calls (citing the work of Paul Dirac) the “choice on the part of nature” (ibid., p.24) – is the statistically generated answer to a “Yes/No” question that ends up being posed (concerning a particular outcome) by the Process 1 intervention. Process 0 is the process that precedes Process 1 and is effectively (on my understanding) the process rooted in an agent’s current established psychology (CEP) – featuring their desires, beliefs, memories and so forth – which determines what the actual Process 1 intervention (the “free choice of the experimenter/observer”) will be in the first place. It is this Process 0 (and, by extension, its resultant Process 1 action) that Stapp claims is not in any way determined by the quantum laws.

> It is the absence from orthodox quantum theory of any description on the workings of process zero that constitutes the causal gap in contemporary orthodox physical theory. It is this 'latitude' offered by the quantum formalism, in connection with the "freedom of experimentation" (Bohr 1958, p.73). that blocks the causal closure of the physical, and thereby releases human actions from the immediate bondage of the physically described aspects of reality. (Stapp 2011, p.24)

Thus, quantum states left to their own devices will continue to evolve, via Process 2, as an ever-growing smear of superposed possible states, representing a host of possible "actual" outcomes but never actually collapsing into any kind of classical reality. A free agent, separately from such a quantum-mechanical evolution (even one happening within their own brain) can – in some largely unspecified way – psychologically decide, via Process 0, to enact a particular choice or decision and then implement a Process 1 action that partitions the quantum state (essentially collapsing the wave function), ultimately selecting a particular outcome that she wishes to become actualised. Whether or not this outcome *will* become actualised is then put to nature to decide (via the statistical Process 3) in the form of a "Yes/No" question. A "Yes" answer means the outcome is indeed actualised, the agent gains an increment of knowledge or has a perceptual experience, and the physical state fully collapses into one congruent with the psychological experience. A "No" answer means the state fails to actualise and the process (presumably) starts over.

There are (again) many questions raised by such a model but, before I begin to consider them, there are two other components of Stapp's theory that need to be introduced: templates for action and the Quantum Zeno Effect.

Templates for Action and the Quantum Zeno Effect

Stapp argues (again, similarly to Penrose and Hameroff) that in order for a freely selected quantum brain state to be able to co-ordinate the required physical activity to bring about an intended behaviour or realise a particular experience, it must be able to endure for a relationally significant period of time. Stapp calls such a state a "template for action." The relevance of this appears to be that after a Process 1 intervention has partitioned the quantum state and posed its "Yes/No" question and, say, a "Yes" result has been delivered by nature (by way of Process 3), the resultant "Yes" state is then positioned as a template for action and needs to be held in place for long enough for the intended behaviour to be realised.

Stapp identifies that, thus far, the only part of his model where the agent can seemingly exert any kind of "free action" is in the selection of which Process 1 action to perform and when to perform it. The actual result of the action would appear to be decided upon by the quantum statistical laws and, even then, there is no guarantee that the resulting template for action will be held in place for long enough to produce the intended behaviour. His answer to this is to introduce what he refers to as the "Quantum Zeno Effect": a feature of quantum theory which he states offers "a way to overcome this problem, and convert the available 'free choices' into effective mental causation." (Stapp 2011 p.35)

Stapp explains the phenomena as follows:

> Suppose a process 1 query that leads to a 'Yes' outcome is followed by a rapid sequence of very similar process 1 queries. That is, suppose a sequence of identical or very similar process 1 actions is performed, that the first outcome is 'Yes', and that the actions in this sequence occur in very rapid succession on the time scale of the evolution of the original 'Yes' state. Then the dynamical rules of quantum theory entail that the sequence of outcomes will, with high probability, be held approximately in place by the rapid succession of process 1 actions, even in the face of very strong physical forces that would, in the absence of this rapid sequence of actions, quickly cause the state to evolve into some very different state. (Stapp 2011, p.35)

The Quantum Zeno Effect is this "holding-in-place" of the relevant state. Stapp, further, equates the *choice* on the part of the agent to implement such Process 1 actions with *mental effort*, meaning that such mental effort can increase the number and rapidity of Process 1 actions which, in turn, can help to generate the required brain activity to bring about the intended action; potentially affording the agent an extra level of control over the ultimate outcome. Stapp sees this mechanism as an example of how mental effort can have a powerful effect on the brain and thus concludes that the quantum laws bestow the potential for efficacious conscious effort onto free agents.

Putting Stapp into Practice

There are a number of immediate issues raised by these additions to Stapp's model but, before I discuss these more fully, I will attempt to extrapolate Stapp's model up to the actual level of specific human action. Stapp himself produces limited "real-world" examples, but one he does discuss briefly is the act of raising one's arm.

> Suppose you are in a situation that calls for you to raise your arm. Associations via stored memories should elicit a brain activity having a component that when active on former occasions resulted in your experiencing your arm rise, and in which the template for arm-raising is active. According to the theory, this component of brain will, if sufficiently strong, cause an associated process 1 action to occur. This process 1 action will partition the quantum state of your brain in such a way that one component, labelled 'Yes' will be this component in which the arm-raising template is active. If the 'Yes' option is selected by nature then you will experience yourself causing your arm to rise, and the state of your brain will be such that the arm-raising template is active. (Stapp 2011, p.35)

Additionally, the employment of the Quantum Zeno Effect would presumably play its part in holding in place the relevant state for the required amount of time to do its work, and thus allowing the agent's "mental effort" (in terms of the repeated Process 1 actions) to play a causal role in bringing about the intended action. One key issue is, however, that even if we were to accept we have this ability to increase the efficaciousness of Process 1 actions, it says nothing about the CEP-based process preceding Process 1 (Process 0) which is responsible for *deciding* what the Process 1 action is designed to achieve – which in this case is to raise an arm.

If we apply Stapp's model to our (more psychologically involved) Jessica example, we might imagine it would proceed as follows. When Jessica witnesses the man fall to the ground, the sensory information generated within her brain presumably forms part of a growing informational/quantum superposition of brain states that extends to include the various possible future decisions she might make and outcomes that could potentially occur. Left to its own devices (without any Process 1 interventions) the quantum state would continue to evolve mechanically without ever realising any *actual* outcome. However, Jessica is capable of deciding via a CEP-based Process 0 (which is not determined or even described by the known quantum laws) what action to take, and then instigating the relevant Process 1 interventions designed to actualise the required physical states to put her desires into practice. Thus, after reaching a decision about what to do (Process 0) – say, to drive on to the most important race of her life and leave the fallen man to his fate – Jessica mentally instigates the relevant Process 1 action which partitions the quantum state of her brain so that a particular "Yes/No" choice is presented to "nature" (to be decided upon subject to the statistical quantum laws involved in Process 3), where a "Yes" answer corresponds to the "driving on to her race" template for action being active and a "No" answer results in nothing at all. If a "Yes" answer is returned, then Jessica can immediately instigate a further series of rapid Process 1 actions which (within the laws of quantum probabilities) could potentially hold in the place the template for action (presumably consisting of quantum states resulting in a relevant pattern of brain activity) and would likely then lead to Jessica actually driving on to her race.

Once again, a number of immediate questions are raised by such an extrapolation of Stapp's model, including the relationship between microscopic and macroscopic brain states, the relevance of quantum effects in the human brain environment,[81] the ontological status and separation (and potential interaction) of the mental and the physical, the existence/relevance of the Quantum Zeno Effect, and many more. However, one of the most important issues for my purposes is one that has arisen in the previous discussions of other libertarian theories: that the actual *decision* for which course of action to follow (Process 0 in this case) – where presumably Stapp (as do others) would wish to ground any notion of ascribing responsibility – comes *before* the most involved parts of the model, and with no discussion for how it is arrived at. As with the Orch Or theory discussed in the previous section (and the Criterial Causation theory in the section before that) the entire focus is on how "mental events" could be causally efficacious over the "physical aspects" of our world, without any discussion of how such mental events themselves are formed or why they may have the content they have.

In addition, there is the issue of the "choice on the part of nature" (Process 3). The answer to the "Yes/No" question posed by the Process 1 intervention seems to be reliant on essentially random quantum probabilities so, even if we do accept that Processes 0 and 1 are in some way under the direct and autonomous control of the agent, there is still a point where what the agent *actually does* is subject to random factors. Stapp's answer to this problem – affording the agent the power to take advantage of the Quantum Zeno Effect in the manner just described – appears only to make *nature's choice* more likely to be implemented. It does nothing to remove the arbitrariness involved in that decision.

[81] Stapp's view on this is similar to Tse's in that he argues that due to the relative sizes involved at neural synapses quantum effects are very much in play. See Stapp 2011, p.30.

3.4.3 Luck, Regress, and the Inevitable Catch-22

The Luck Objection

The crux of the Luck Objection, it will be recalled, is that for an agent to be truly free (in the Gold Standard libertarian sense) she must be able to both ‘do’ and ‘do otherwise,’ all past circumstances remaining exactly the same (must meet the AP condition); but if there can be nothing in her past (including her own psychological make-up, her CEP) that can fully determine her decision or ultimate action, then what *is* decided can only be down to luck. If an agent such as Jessica with *exactly* the same past, with *exactly* the same CEP, and after *exactly* the same deliberations, could still just as easily follow either course of action – what *else* but luck could be the deciding factor? Though he does not directly address the Luck Objection or feature much discussion on the concept of luck in human agency, I conjecture that Stapp’s position would be that he has afforded the agent sufficient control over her choices and actions so that such choices would not be considered arbitrary. He has argued that the complete freedom of the agent to instigate undetermined Process 1 actions allows her to have direct causal control over her choices, and in such a way that the choices are not accounted for in (and thus not determined by) the known quantum laws. He has also argued that the employment by agents of the Quantum Zeno Effect gives them even greater control over whether their chosen course of action will ultimately be implemented. But the question is whether either of these factors would satisfy a proponent of the Luck Objection.

By Stapp’s own admission, the “choice on the part of nature” (Process 3) is essentially random – being subject to the statistical laws of quantum probabilities – and, as we have just discussed, the addition of the Quantum Zeno Effect seems merely to make whatever choice *nature* makes more likely to actually occur. In addition, with no discussion of the ontological nature and structure of the important Process 0 – the (presumably conscious) CEP-based decision making process that determines the choice and timing of the Process 1 intervention it precedes – essential, decision-making thoughts would appear, if they are not to be subject to an infinite regress, to arise non-deterministically from within the agent’s CEP further strengthening the charges of luck and arbitrariness due to the apparent causal disconnection. It could well be argued that, on Stapp’s model, *all we in fact have* is a collection of random events: from the Process 0 decision of how to act arising non-deterministically (and thus arguably arbitrarily), followed (after the relevant Process 1 action) by an essentially random statistical Process 3 as to whether or not the quantum laws return a result compatible with carrying out the decided upon act, through to a situation where a template for action relating to enacting the act may or may not be held in place for long enough for the relevant brain processes to take hold.

Stapp actually readily acknowledges the involvement of random factors, as he sees them as essential to distance his quantum model from what he considers to be the completely determined, mechanistic and free-will-lacking picture created by competing models developed on a foundation of classical physics. However, he sees such randomness limited solely to the Process 3 element of his model and – because of agent-influencing factors before and after this part – does not appear to consider such randomness problematic: quite the opposite.

> The inclusion of the quantum element of random chance can rescue the meaningfulness of one's life. For these choices are not written in stone, or fixed by mechanical determinism. Hence, by being fundamentally indeterminate, or "random", they can be become biased by values, which can thereby influence the course of physical events! (Stapp 2017, p.38)

Stapp's use of such phrases as "biased by values" – as well as his belief in the general efficaciousness of mental events – implies that these essential indeterminacies apply exclusively on the physical side of the equation. The physical aspect of reality (as explained in the mathematics and governed by the known quantum laws) must be indeterminate in order for mental events (and consciousness generally) to have any influence and restore the free will that was eradicated by the solely physical, deterministic and causally closed classical picture, but no consideration is given to the implications of the widespread application of indeterminism to the mental events themselves (or, for that matter, to how we might understand free actions within a deterministic setting). As we have seen in previous chapters, even if Stapp's model did not have points post Process 1 where random factors are influential, the target point for the Luck Objection would just shift back to Process 0 and the conscious formulation of the initial choice from within the agent's CEP. If that choice remains undetermined, as Stapp would likely need it to be, the problem of luck does not go away. And if it *is* determined, we face the inevitable regress, to which I will now turn.

The Basic Argument

As in previous sections, our other main objection focusses on the regress just discussed; a regress seemingly (and inevitably) created by any attempt to completely remove luck and randomness from the process of decision-making and action. This problem is highlighted by the Basic Argument which (as a brief reminder) states that we can only be held responsible for our decisions and actions if they flow from our psychological states (from our CEP) but, be that the case, we face a vicious regress of decisions arising from a CEP which was formed from previous decisions made by a previous EP and so on back to a time when we would not have been capable of making any attributable decisions for ourselves.

Again, although Stapp does not consider this argument directly, it is safe to infer that he believes his model to be sufficient for the attribution to human agents of this kind of responsibility.[82] Stapp is clear that he sees the Process 1 interventions – the metaphorical "free choice of the experimenter" – as key to breaking any regress he argues is inescapable on the classical picture, so would likely highlight that process as a point in his model which could offer a challenge to Strawson's argument. There is also Process 3 – nature's statistical choice – which is, in an important sense, causally disconnected from the antecedent conditions as it is essentially a random outcome based on the statistical laws of quantum probabilities. Further, even *after* nature's choice is returned one way or the other, whether or not the resultant template for action remains in place for a sufficient length of time to be effective is also indeterminate, adding yet another point where events can be causally disconnected from the antecedent conditions.

Highlighting such points at which a regress could be halted raises two main issues. The first is that (as with the other models discussed up to this point) Stapp undoubtedly wants to have his agent's choices satisfactorily rooted in their CEP, which would emphasise the importance of the Process 0. This process presumably includes the relevant deliberations and incorporates the relevant desires, beliefs, past experiences and so forth. The resultant decision which then determines the Process 1 intervention is

[82] As an example: "Our legal system is based on the idea of personal responsibility for one's physical actions. But, according to classical mechanics, every physical action was predetermined at the birth of the universe. A person cannot rationally be held responsible for physical actions that were physically pre-ordained at the birth of the universe. Quantum mechanics, on the other hand, does not entail any such physical predetermination, and thereby evades the classical-mechanics-based challenge to the rationality of our justice system!" (Stapp 2017, p.119)

generated mechanically from such psychological activity so (regardless of whether this process takes part in the physical brain or in a dualistic mental realm)[83] presumably the CEP which generates and instigates the reached upon decision must have been formed by past decisions, experiences, and psychological states. The regress, therefore, is not eradicated but simply shifted back a stage to Process 0. If Process 0 is not ultimately determining of what Process 1 action occurs then, as we have seen previously, we are back to the problem of luck; and the same would be said if we try to insist that the regress can be halted at the Process 3 point and subsequent template-for-action stage. Ultimately, any attempt Stapp might make to show how his model might tackle Strawson's argument leads him right into the *Catch-22 of Libertarianism*: if he tries to suggest the regress can be halted by imposing any kind of causal disconnect between Process 0 and Process 1 (or somewhere around process 3), he becomes vulnerable to charges of luck, and if he tries to ensure full agent control by having Process 1 actions effectively determined by Process 0 (and somehow carrying this determination all the way through the subsequent processes) the regress simply shifts to whatever processes precede Process 0. Once again, the *Catch-22* appears inescapable.

3.4.4 The *Criteria*

I will now review Stapp's model against my *Criteria for Coherence*, beginning once again with the first two principles taken together.

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.
P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.

Stapp is clear throughout his work that he believes a libertarian free will does exist and that it is incompatible with classical mechanics *because* such a model is entirely deterministic. Stapp also emphasises the importance of quantum indeterminism as key to breaking what he considers to be the "causal closure" of the physical universe, as it is described by classical physics. As his theory is based

[83] It is not completely clear whether Stapp does ultimately support a fully dualist ontology (in the immaterial sense) or whether it is more the case that the psychological processes involved are rooted in a different kind of physical activity, not necessarily described by the quantum laws that govern the other aspects of physical reality. See Stapp 2011, chapter 12 – Despised Dualism – for more on this issue.

entirely on his preferred interpretation of quantum mechanics and draws heavily on the indeterministic nature of the quantum laws, I believe we can safely say that both principle one and two have been met.

P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.

Though he does not appear to directly discuss this notion of "doing otherwise" (or meeting the AP condition generally), Stapp would no doubt point to his highlighted concept of the "free choice on the part of the experimenter" and that, far from being deterministically governed by the quantum laws, such choices are not even included in the mathematical descriptions at all. If it is true that there are such inescapable gaps in the quantum dynamics, and that mental events of an autonomous agent can fill these gaps in such a way that is entirely unspecified by the physical laws, then it seems clear the choices of that agent are in no way determined by such laws and the agent is indeed free to act or act otherwise, whatever the past conditions and natural laws happen to be. Thus, this principle is also met.

P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.

As ever, here is where the theory begins to fall down as a libertarian account of free will. As we have seen above, Stapp's main attempt to avoid determinism is to emphasise the indeterministic and "psychophysical" nature of quantum mechanics, which he believes to provide the true description of our universe. Within this structure, free agents are able to take advantage of the fact that relevant events have both a psychologically described aspect and a physically described aspect, and that the mathematical descriptions of the quantum laws only account for the physically described aspect (which, in addition, is incomplete), leaving a causal gap that mental events can exploit in order to afford full control to an agent. This picture, Stapp believes, frees human agents from the chains of a classical, deterministic picture in which their thoughts and desires can have no efficacious effect (presumably because they are either reduced out of existence or rendered epiphenomenal). Moreover, by inserting the importance of conscious choices into the dynamics and having them become, not only causally efficacious, but also essential in order to resolve superposed quantum states, Stapp believes he is providing a model that does confer sufficient control onto an agent in an indeterministic universe.

However, the introduction of a consciousness that is in some way independent of the known physical laws does not, of itself, do anything to resolve the problems that have been highlighted in regards to determinism. All that occurs, is that the form such determinism takes changes from the standard physical determinism Stapp associates with classical physics to a different type of psychological regress associated with Process 0's determined influence over the Process 1 interventions. In addition, the Process 3 "choice on the part of nature" means that whatever decision *is* actually decided upon may or may not be put into practice depending on random quantum probabilities. Stapp's efforts to employ the Quantum Zeno Effect to remove some of the randomness and bolster the control of the agent do little to help this issue, as all such repeated Process 1 actions seem to be able to do is to hold in place (and therefore make more likely to occur) whatever choice "nature" has already (arbitrarily) made.

Ultimately, on Stapp's model, the decision of what action the agent wishes to try and implement (via Process 1 and 3) appears essentially determined by prior psychology (culminating in Process 0) and whether or not the chosen action is actually implemented is largely arbitrary, thus seemingly offering the worst of both worlds. Thus, principle four cannot be said to have been met.

P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.

As above, Stapp's model fails to remove any regress, because the main locus of control appears just to be shifted to an independent consciousness which still remains an established psychology; presumably comprised of thoughts, beliefs, and desires (and so forth) which are born out of past learning and experience. Attempting to halt any regress by claiming there is a causal gap in the physical dynamics which requires the intrusion of mental events from such a consciousness just shifts the regress from the "physical" to the "mental" (whatever we take these terms to mean). In any case, the causal effectiveness of Stapp's agent is severely hampered by an apparent failure to be able guarantee that any of the chosen desires will be put into practice. Principle five has also, therefore, not been met.

P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

Whilst Stapp would undoubtedly argue his theory is based on solid science and empirical findings, there is much in his model that would likely inspire a good deal of debate. In particular, the kind of dualism Stapp seems to be postulating is somewhat mysterious; both in relation to the ontological status of mental events and the process of interaction with physical ones. Stapp's psychophysical events are also somewhat hard to comprehend with the notion of physical states only becoming *real* in order to correspond with a psychological entry into the stream of consciousness of a human agent raising many questions regarding the workings of our universe prior to the emergence of consciousness. Ultimately, Stapp's position certainly appears somewhat obscure and, given its failure to address the key problems I have raised for libertarian free will, the price of acceptance would seem unnecessary to pay.

3.4.5 Conclusion

In the end, Stapp's theory appears to be more focused on rationalising how mental events, such as thoughts and feelings (which appear in some way independent of the standard physical laws), could enter into the physical dynamics; indeed, are *required* by the physical dynamics due to the evolution of a quantum state being *unable* to resolve itself in such a way that would ever correspond to a recognisable singular state. It could perhaps be argued that Stapp is guilty of taking the phrase "free choice of the experimenter" that was introduced by the founders of quantum theory rather *too literally*, and the phrase was possibly not designed to imply that some kind of independent consciousness (entirely separate from the mathematically described quantum laws) is needed in order to resolve an evolving quantum superposition. Rather, it could simply be a statement of the influence of the experimenter's decision of what experiment to perform and how/when to go about performing it, leaving the mind of the experimenter and its physical substrate firmly withing the same scientific structure as the planned experiment and the quantum laws guiding it.

In any case, as I have stated above, even if Stapp's model (and its implied psychophysical picture of the universe) were to be entirely accepted, it has ultimately failed to meet the *Criteria*. Far from succeeding in removing all traces of randomness or arbitrariness *and* halting the problematic regress, Stapp's model for the Gold Standard free will actually manages to do neither and, thus, like all the other theories considered in this thesis, cannot be considered fully coherent.

3.5 Where To Now?

In Part Two of this thesis, I have considered three theories developed by scientists and evaluated these against my *Criteria for Coherence* to see if they can meet the standards that I (and they) believe must be met in order for their imagined agents to achieve the Gold Standard freedom of the will. I have argued that (as in Part One) none of the theories considered has satisfactorily met the principles contained within the *Criteria* and thus cannot, by these standards, be considered coherent.

Peter Tse offers a model for libertarian free will which focusses on the active role played by neurons in the human brain; neurons that can recode their own response to stimuli and alter the informational criteria for when they will or will not fire. When this neuronal activity is combined with inputs that vary indeterministically and a coordinating consciousness (he argues), an agent can be capable of non-determined but self-selected actions. However, Tse fails to offer a satisfactory explanation for how the relevant decisions are made in the first place and his model likely includes *both* determined and random processes, rather that offer any "middle road" between them.

Roger Penrose and Stuart Hameroff attempt to show where, within the physical workings of the human brain, consciousness might be created and also how it might exert its influence over the agent's physical actions. By "orchestrating" quantum activity and state-reduction events within neuronal microtubules, agents can create and employ an efficacious consciousness in a way that is neither completely determined nor completely random. Precisely how such a faculty works is, however, quite vague and their model also does not manage to fully avoid either randomness or a problematic regress, leaving it vulnerable to both of the key objections.

Henry Stapp's quantum mechanical approach requires an intervention by an observer into the basic physical dynamics in order to collapse the quantum wavefunction and resolve the evolving superposition of possible states. Such an intervention is not determined by anything described in the known quantum laws and thus can attribute sufficient control to an agent. However, the undetermined acts of the agent are not all that is required in decision-making and action and much of his model is reliant on statistical processes over which the agent can seemingly have little or no control. As such, Stapp (like the others) fails to avoid some level of arbitrariness or some form of regress and, furthermore, his theory relies on a number of unexplained and arguably mysterious interventions.

With none of theories from either Part One or Part Two adequately meeting the *Criteria* – and thus not being able to be considered coherent – we seem to have failed to find a libertarian theory which can satisfactorily and coherently provide a model for action and decision-making which meets all of the libertarian goals and thus achieves the Gold Standard freedom of the will we have been pursuing. If such an ideal is potentially *impossible* to realise in any intelligible fashion, the question we must then consider is: what it is about this Gold Standard that makes it so difficult to attain?

Conclusion

The title of this thesis poses an important question: Is libertarian free will an inescapably incoherent concept? In order to provide an answer, I have proposed a *Criteria for Coherence* against which the featured libertarian theories have all now been assessed to see if any can adequately meet all its principles. As discussed in the introduction, the principles contained in the *Criteria* are derived from the key tenets of the libertarian approach – as well as from considerations which arise out of libertarian attempts to defend the position against key objections such as the Luck Objection and the Basic Argument – and fully reflect the stated aims of the researchers whose theories and models I have considered.

The *Criteria for Coherence* proceeds as follows:

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.
P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.
P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.
P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.
P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.
P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

Theories from philosophers representing the three main libertarian positions (event-causal, non-causal and agent-causal) and three theories developed by scientists who have worked on the problem (Criterial Causation, Orchestrated Objective Reduction and Quantum Interactive Dualism) have all been evaluated against the *Criteria* and all have failed to satisfactorily meet its principles, meaning none of them (by these standards) can be considered coherent. I shall now briefly summarise each approach and where it falls down, before moving on to discuss the implications of this wholesale failure.

4.1 Theories Vs the *Criteria*: A Summary

4.1.1 Robert Kane's Event-Causal Theory

Robert Kane's event-causal model is a well-developed and intuitively appealing attempt to present an intelligible libertarian theory. Kane introduces his "self-forming actions" (SFAs) within a framework of events internal to the agent causing (albeit sometimes non-deterministically) other such events in a broadly straightforward and linear fashion, which can ultimately lead to decisions and actions. These key moments in the life history of the agent are said to be self-generated in a way that is "screened off" from the past and are thus undetermined and regress-stopping and their incorporation into Kane's model allows it to meet both the alternate possibility (AP) and ultimate responsibility (UR) conditions; endowing his agents with the elusive antecedently undetermined – *yet still determining* – control. With such SFAs forming the foundation of an agent's will (or CEP), the less important day-to-day decisions can flow from it in an essentially determined fashion without this impinging on the agent's libertarian freedom. The core "will" is formed and maintained by the SFAs and thus *all* the agent's actions can be considered truly free (to the Gold Standard level), so long as they remain rooted within such a will. A satisfactory form of undetermined control can thus be attained by the agent on such a picture because the agent is succeeding in achieving what they were trying to achieve – *whichever of the desired outcomes is ultimately realised* – and they subsequently consciously recognise the ultimate action as their own; as what they were trying to do. It does not matter that their psyche may have been trying to achieve various different contradictory courses of action simultaneously. By achieving what they were trying to achieve, Kane argues, we would never call the result just a matter of "luck" or "chance" or label it as being an "arbitrary" outcome and the dismissal of such notions (along with the absence of deterministic outcomes) just is what "control" *is*, in this context.

For all its positives, however, the theory fails to meet a number of the principles I have set down in order to judge its coherence, and thus it fails also to fully meet its author's own stated goals. Whilst we may grant (though not without some reservations) that Kane's theory adheres to the first three principles and principle six, I have argued that he has not managed to provide an account of undetermined decision-making that would not be considered fundamentally arbitrary in nature. Or, to put it another way, he fails to adequately conjoin undetermined decision-making with a sufficient level of agential control in such a way as to permit any justifiable ascription of responsibility (falling foul of the ever-problematic *Catch-22 of Libertarianism*) and there remains no understandable reason or existing physical factor for why – during an SFA – one course of action becomes favoured over another and is thus ultimately enacted. Kane's theory, therefore, fails to meet principle four. With regards to principle five, some form of regress appears inevitable despite Kane's attempts to eradicate it as, even if we accept the final decision during an SFA to be undetermined, there is the matter of other parts of the decision-making process (such as the arising of desired options) arguably manifesting in a determined fashion.

In summary, Kane has not managed to provide a model for free action and decision-making that can be said to avoid all forms of regress *and* be fully within the agent's control. This is because his attempts to remove the regress inevitably result in an unacceptable, responsibility-reducing level of arbitrariness being present in the system, and his attempts to remove the arbitrariness arguably just re-introduces some form of problematic regress.

Verdict: Fails to meet the Criteria

4.1.2 Carl Ginet's Non-Causal Theory

Carl Ginet's approach follows the non-causal model by arguing that free actions need not be *caused* by anything at all, including the relevant antecedent states of the agent (such as beliefs, desires and so forth). Removing any causal link, Ginet believes, takes with it any problematic regress and allows the agent to pursue undetermined courses of action. Actions, Ginet argues, start with a basic, causally simple mental event (such as a volition), which is itself not caused by any antecedent mental events, and what makes such mental events an *action* at all (and not just a random occurrence) is an accompanying "actish" quality. It is this phenomenal quality that makes the agent the subject of her action: it is what makes it seem to the agent that it is *her* action and not just something that happens to her. What then makes such an action *free* is the fact that it is not causally necessitated by antecedent events. Thus, the way the agent has control over such actions is by *feeling* (phenomenologically speaking) that they are in control of them – that it is *them* that is performing the action – and by still being able to explain a chosen action in terms of the reason or purpose for carrying it out (via a referential concurrent intention) despite the broken causal link. It is in such a way that Ginet believes he has afforded his agent the undetermined/non-random Gold Standard free will he, like his libertarian colleagues, sets out to achieve.

When considered in detail, however, it is difficult to see how Ginet's theory avoids *either* determinism *or* arbitrariness. For example, there is no discussion concerning how the guiding desires are formed but if they are not to be completely random arisings they must presumably be determined by the agent's CEP – which instigates a regress in that part of the process. In addition, these desires are then (in Ginet's attempt to halt the regress where he believes it is most problematic) apparently completely causally disconnected from the subsequent action. We appear to be left, therefore, in a "worst of both worlds" situation where guiding desires arise in a determined fashion but do not determine the outcome, leaving the agent in the strange position of not being able to control whether or not the outcome favoured by the (deterministically produced) guiding desire is actually implemented. Far from providing a model for decision making which would be considered neither determined nor arbitrary, therefore, Ginet has managed to produce one which seems to be both, placing his theory in the position of being subject to versions of both the Luck Objection and the Basic Argument.

Whilst principle five may arguably appear question begging from the outset – as in Ginet's theory the agent is not the *cause* of anything at all – Ginet is still attempting to escape a problematic regress and his theory fails to achieve this because, as just discussed, the regress simply shifts from the disconnected action to the prior reasons/desires guiding the action. Principle six provides further reason not to accept non-causal theories such as Ginet's because, whilst it could perhaps be argued that removing the requirement for any causal connection removes any need to develop obscure forms of the causal relationship (of which both event-causal and, in particular, agent-causal theories have been accused of doing), the result here of such a move is a theory which just appears overly obscure generally, so we do not seem to be in any better position. The causal picture is at least usually comprehensible, and with no satisfactory positive account of the ontological relationship between an agent's CEP and her actions we cannot say principle six has been met.

In summary, in addition to being somewhat obscure and generally rather counter-intuitive, Ginet's non-causal account fails to solve any of the problems facing libertarian theories and, again, offers no satisfactory model that combines undetermined decision-making with sufficient agential control. It is not in fact clear that his model offers either.

Verdict: Fails to meet the Criteria.

4.1.3 Timothy O'Connor's Agent-Causal Theory

Timothy O'Connor's agent-causal approach attempts to provide the agent with an enhanced form of control that O'Connor believes is missing from both the event-causal and non-causal theories. The agent – as an enduring substance irreducible to events – has full, free, and direct control over their decisions and actions by way of a special, primitive causal relation (also not reducible to causation by events) which allows them to directly cause their actions or intentions to act. Agents are true *prime movers unmoved* – the creators/initiators of new causal chains who cannot themselves be caused to create such chains by anything else. By exerting an "active power" to bring about specific intentions (which is not considered an event and which is not itself caused and so requires no prior act of will), O'Connor believes agents on his account can avoid being subject to either an unacceptable level of arbitrariness or an infinite regress, and thus can attain the Gold Standard freedom of the will.

O'Connor defends his theory by arguing for the acceptance of an irreducible, non-Humean causation (whereby an agent can freely exercise a different type of causal power or capacity that is distinct from, but parallel to, his non-reductive event causes) and also attempts to provide an account of reasons-explanation of free action, as agent-causal theories suffer from similar problems to non-causal theories in this regard. Here, O'Connor adds his agent-causal capacity on top of Ginet's model, which then acts on the arising desire by causing the "coming-to-be-of-an-action-triggering-intention-to-act" in such a way as to satisfy this desire. The desire itself does not directly cause (and certainly does not necessitate) the action but is continued to be held throughout to sustain the action, which is caused directly by the agent.

However, due to its similarities to Ginet's view, O'Connor's account of reasons-explanation is susceptible to all the same criticisms; including the difficult issue of the arising desire being either determined to occur or formed arbitrarily, plus the absence of any easily understandable connection between this desire and the act itself. In addition, O'Connor's approach requires a somewhat mysterious agent-causal capacity for which we lack a clear description: either of what it *is*, or of how it *works*. Similarly to the event-causal and non-causal theories, principle four provides the first main stumbling block as, once again, it is both unclear how the important guiding desires/concurrent intentions are formed *and* unclear how they are connected to the subsequent decisions or actions. It is therefore difficult to see how O'Connor can claim to have escaped *either* determinism *or* arbitrariness, and thus unclear how the agent can exert the superior level of control O'Connor is claiming she can wield. Also, considering principle five, we cannot say for sure that the regress has been avoided either, as it appears simply to have been shifted to the formation of the guiding desires and/or the implementation of his "structuring causes," the latter of which, he argues, likely inclines an agent to one outcome over another.

O'Connor's theory is also arguably the furthest of all the libertarian approaches from meeting the sixth and final principle. To many, his requirement for an enduring (and ill-defined) agent-as-substance, his reliance on a non-Humean (non-reductive) causation, and his "causal powers" concept of causal production all entail arguably unusual ontological commitments they are not inclined to accept – particularly because O'Connor's model does not seem to solve any of the problems that beset the libertarian position.

Verdict: Fails to meet the Criteria.

4.1.4 Peter Tse's Criterial Causation Theory

Peter Ulric Tse attempts to develop a theory of free will from his neuroscientific research and argues that it is indeed possible to have a libertarian model that allows for undetermined but entirely self-selected actions (the Gold Standard). Key to his theory is that neurons, in his view, are not simply blunt, stimulus-response devices solely at the mercy of their presynaptic inputs but have the ability to recode their own response to stimuli – and that of other neurons around them – altering the informational criteria for when they will or will not fire. Patterns of neuronal activity harnessing this ability can therefore be causal in a way over and above the physical basis in which such patterns are ultimately realised and it is by utilising such a biophysical mechanism that an agent can exert a form of top-down mental causation. When we add indeterminism into the picture so that the neuronal inputs received are inherently variable and unpredictable and combine this with an agent's ability to consciously manipulate the contents of our endogenous attentional operations, we have (according to Tse) all the ingredients we need for what he terms a “strong free will.”

As with the other approaches, Tse’s theory falls down when he tries to develop a model for precisely how this self-selecting but undetermined process of enhanced agent control would work. It is plausible to think that Tse himself would argue that principle four has been met, but all the control he offers through criterial encoding occurs *after* the key conscious decision has been made. There is no discussion concerning exactly how the decision/desire is arrived at in the first place, leaving his view being subject to the familiar objections (in lieu of an account to the contrary). With prior desires and so forth being, once again, either deterministically generated or arising randomly out of our CEP, we are no further along in our attempts to escape from the *Catch-22 of Libertarianism* and whichever way Tse decides to view the desire generation he will either fall foul of the Luck Objection or the Basic Argument. Plus, the inclusion of indeterminism by way of random neuronal inputs arriving at (consciously) preconfigured synapses brings into question precisely *how much* control Tse’s model is providing the agent in any case.

Ultimately Tse seems more to be presenting a useful neurological mechanism for how an agent can act upon a deterministically generated intention and, as such, is arguably developing a model more for freedom of action than for the Gold Standard freedom of the will he is claiming to have achieved.

Verdict: Fails to meet the Criteria.

<u>4.1.5 Roger Penrose and Stuart Hameroff’s Orchestrated Objective Reduction Theory</u>

The Penrose-Hameroff Orch OR theory seeks to propose – and take advantage of – a gap in the apparently computational workings of the physical human brain, right down at the quantum level where consciousness is not only created but where it can freely exert influence over physical matter, and thus control bodily behaviour. It postulates that quantum activity within microtubules creates a vast network of entangled particles (via gap junctions) in superposed quantum states, which evolves deterministically in line with the Schrödinger equation until the threshold for Penrose’s (non-computational, gravitational) Objective Reduction is reached. At this point, the wavefunction collapses, a new and unknown process takes place, and a moment of “proto-consciousness” occurs. By “orchestrating” such quantum activity and state-reduction events, agents can (presumably in conjunction with their CEP) create and employ an efficacious consciousness – influencing neuronal firings/inhibitings that can presumably control actions. The backward referral in time of quantum information is proposed to counter any arguments that the created consciousness is solely epiphenomenal and to try and prevent the model being viewed as either a completely deterministic or a completely random process, thus establishing a Gold Standard of free will for the agent that is neither determined nor arbitrary.

What is not discussed at any length, however, is exactly *how* such orchestration could produce any kind of coherent psychology that would presumably be making the decisions that need to be enacted via the Orch OR neuronal mechanism. As presented, the output of the quantum-conscious decision-making process would appear arbitrary – at the whim of quantum probability – and the theory does not in any way explain how a model for such decision-making could be undetermined by past events but remain under the complete control of the agent. As with Tse's model, there is a mechanism for how a conscious will might causally affect indeterministic neuronal processes (though not as mechanically specific as Tse's), but no real explanation for how/why said conscious will comes to decide what course of action it wishes to enact in the first place. As the theory appears to come down more on the side of quantum indeterminism than determined decision-making it is particularly susceptible to the Luck Objection; indeed, the whole nature of consciousness generation and any derived decision-making process would appear solely arbitrary. However, any attempt to root the conscious decisions in a coherent psychology would then only reinstate the familiar regress.

Given the (acknowledged) number of unknown elements in the Orch OR theory, the whole approach is arguably somewhat mysterious. The central process that occurs at the point of wavefunction collapse and generates consciousness is entirely unknown. There is also no real explanation for what a "proto-conscious moment" actually is, how a collection of such moments could come together to form a coherent subjective experience, or how conscious decisions are generated in this process. Penrose and Hameroff might wish to argue these particular mysteries are at least scientific mysteries, but mysteries they remain so, as well as principle four and five, principle six is also not met.

Verdict: Fails to meet the Criteria.

4.1.6 Henry Stapp's Quantum Interactive Dualism

Henry Stapp's quantum mechanical approach is born out of his firm belief that a universe beholden to classical physics would entail epiphenomenalism (if not eliminative materialism); ruling out consciousness and therefore any form of genuine free will (for which, in his view, consciousness is essential). Stapp subscribes to a variant of the Copenhagen Interpretation of quantum physics which, he argues, does not just allow for the existence of consciousness, but in fact often (if perhaps not always) *requires* its intervention into the basic physical dynamics in order to collapse the quantum wavefunction. In addition, he believes such conscious interventions are themselves completely undetermined and not even accounted for in the known quantum laws. On Stapp's model, the deterministic evolution of the quantum state can be interrupted only by an intervening probing action – which can be carried out freely by a conscious entity – that produces a set of specific alternative possible states that could manifest after the collapse of the quantum-superposed state. After such partitioning, it is "nature" which then has the final say on the outcome, as one of the possible states is selected for manifestation based on the quantum statistical rules (though this choice on the part of nature can perhaps be *influenced* by some form of mental effort employing the Quantum Zeno Effect).

As with all the other theories considered, Stapp fails to provide an adequate model of decision-making and agent control that can completely avoid either arbitrariness or a deterministic regress, thus failing to meet principles four and five. On his approach, there is no explanation for the formation of the (potentially dualistic) conscious will, only consideration for how such a will could (and often needs to) involve itself in the quantum dynamics. This will, then, is once again either subject to a regress or arbitrarily formed. In addition, even were we to ignore the issue of the formation of this will, its only role is apparently to pose appropriate questions (probings) and then await nature's answer, which would appear to significantly weaken the notion of free will Stapp is proposing. And, as with Orch OR, there is a good deal of mystery and obscurity in the model, so principle six is also not met.

Verdict: Fails to meet the Criteria.

4.2 An Impossible “Gold Standard”?

The purpose of applying my *Criteria for Coherence* to the theories considered in this thesis was to test whether any of them have managed to successfully provide a model for free action which can meet all the true ideals of the libertarian position in a coherent manner. I have argued that the type of freedom the authors of the models are trying to achieve is what I have called the “Gold Standard” freedom of the will, which involves (at least some) actions and decisions that are undetermined by antecedent conditions, but which are nonetheless fully within the (essentially determining) control of the agent. Critics have argued that the two conditions are mutually exclusive and therefore successfully meeting one can only lead to a failure to meet the other, and thus libertarian theories that attempt to achieve this goal are inevitably going to descend into incoherence. However, as we have seen, this has not stopped philosophers and scientists with libertarian inclinations from developing models which do attempt to achieve such a Gold Standard.

Ultimately, I have argued that none of the theories I have assessed manages adequately to meet all six principles contained within the *Criteria* and thus none, by this measure, can be considered coherent. What now needs to be considered is exactly *why* they have failed, and what are the implications of this failure.

4.2.1 The *Catch-22* and the Root of the Problem

With the *Criteria* seemingly very difficult (perhaps even impossible) to meet, the natural next step is to turn our gaze inward and consider what, precisely, is it about certain of these principles that makes adhering to them so problematic. In doing so it becomes apparent that the root cause of the failure of all the considered approaches is, in fact, the very thing that makes them libertarian theories – the sacrosanct requirement to include indeterminism explicitly in their models.

Each of the philosophers and scientists whose work I have considered seem quite clearly to assume that the only way principles 3, 4 and 5 can be met is by basing their models within an indeterminist framework (hence the addition of principles 1 and 2), and this assumption can ultimately be traced to the belief that neither the Alternative Possibility (AP) condition nor the Ultimate Responsibility (UR) condition can be met in a deterministic universe. This then leaves them with the difficult (again, perhaps impossible) task highlighted by principles 4 and 5 of needing to develop a model for action which manages to conjoin two necessary, but seemingly contradictory faculties: undetermined decision-making and a sufficiently determining agential control. What we have seen, in both Part One and Part Two of this thesis, is that none of the hugely varying models presented manages to achieve this in any coherent fashion. As the *Catch-22 of Libertarianism* highlights, whenever a theorist tries to evade the Luck Objection and introduce some kind of determining control to reduce the level of apparent arbitrariness in their model (which arises due to this necessary adherence to the AP condition and the indeterminism libertarians are convinced such adherence entails) they increase the likelihood of a vicious regress being instigated, and whenever they try to evade the Basic Argument and introduce more causal breakages in order to reduce the danger of a regress (and thus adhere to the UR condition, for which libertarians again believe indeterminism is essential) they inevitably increase the level of arbitrariness. Often, in such attempts to rid their theories of all arbitrariness *and* a problematic regress, they just end up with models in which their agents are overly subject to *both*.

The question we need then ask is "why should this be the case?", and one answer could be that achieving the level of agent-control libertarians themselves claim they require in order to attain their Gold Standard freedom of the will is just *not possible* whilst adhering to the main (indeterminist) tenets of their own libertarian position. Judging from the wide variety of theories I have considered in this thesis, it does not seem to make any difference whether the theory is grounded in philosophy or science; there just seems to be no libertarian form of control which can match the arguably more straightforward, antecedently determining control afforded to such alternative (deterministic) philosophical positions as compatibilism. Again, we are left to consider why this is so, and the answer strongly suggesting itself is that the indeterminism so categorically required to be featured explicitly in the libertarian models – in order to meet the AP and UR conditions – can only ever confuse any picture of agent-control by emphasising seemingly intractable issues concerning the nature of the link between and agent and her actions; a link we intuitively require an intelligible explanation of in order to attribute a level of agential control that could be considered sufficient for any ascription of responsibility.

Given the above, the *really* important question I believe we should therefore be asking is: is this absolute focus on indeterminism *required* in order to achieve an acceptable model of free action that meets the stated goals of the libertarian project – and thus offers a sufficient level of freedom – and within which an agent can be said to have full control and be responsible for their deeds? To answer this, we must first take a closer look at the two key conditions that appear to be the main driving force behind this focus on the need for indeterminism.

4.2.2 AP and UR

The Alternative Possibilities (AP) condition and the Ultimate Responsibility (UR) condition – regardless of which should actually be considered as the *more* important in debates surrounding freedom and responsibility – do indeed both seem to be behind much of the libertarian impetus to feature indeterminism so prominently in their models for free and responsible actions. This is because libertarian theorists usually consider that it would not be possible to meet either of these conditions should the universe prove to be fundamentally deterministic. Having identified this focus on indeterminism as the main reason behind the libertarian incoherence problem, consideration thus needs to be given as to whether this is actually the case, and whether indeterminism really *is* required to meet either, or both, of AP and UR.

Alternative Possibilities

As just discussed, libertarians generally consider that it is only on an indeterminist account that AP – and its extension into Robert Kane's Indeterminist Condition: "the agent should be able to act and act otherwise (choose different possible futures), *given the same past circumstances and laws of nature*" (Kane 2005, p.38) – can be met, and *not* meeting it would exclude any theory from being able to achieve the Gold Standard freedom of the will due to it inevitably lacking the "deep openness" believed to be an essential requirement for being able to "do otherwise" than what one actually does. As we have seen, however, all the attempts by the models discussed to comply with this condition have ultimately ended up in an incoherent position and so it seems sensible to question whether indeterminism really is essential for a successful adherence.

A number of well-discussed arguments have been put forward which contend that meeting the AP condition can indeed still be achieved should the universe prove deterministic and thus the "deep openness" libertarians claim is essential is *not*, in fact, needed in order for an agent to have genuine alternative possibilities available to them and for them to be able to do otherwise. These arguments have often arisen as specific rebuttals to one of the most widely discussed arguments put forward for the incompatibility of freedom and determinism – the Consequence Argument – which Peter van Inwagen (one of its key proponents) summarises as follows:

> If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born; and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us. (van Inwagen 1983, p.16)

The implications of the Consequence Argument are that for Jessica to be able to do anything other than what she actually *did* do (for her to have done otherwise), she would have to have been able, at the moment of decision, to either alter the physical state of the world at some point in the past (perhaps even prior to her birth) or alter the very laws of nature. As it seems intuitively obvious that she can do neither, the conclusion is proffered that – as our present actions are the necessary consequence of the conjunction of the past and the laws of nature – there is no way for us to change the fact that our present actions occur: there is no way for Jessica to do otherwise than she actually does (or for anyone else to do so, ever). The Consequence Argument is an intuitively powerful argument for the incompatibility of freedom and determinism which underpins much of the libertarian position, but it has been challenged by compatibilists, including David Lewis, who argues that its intuitive appeal is actually based on the conflation of two very different claims about the necessary abilities of free agents.[84]

In his 1981 essay, "Are we Free to Break the Laws", Lewis argues that being able to do otherwise than one actually did in a deterministic universe does not, necessarily, entail the ability to do something as incredible as violate the laws of nature. Lewis distinguishes between a weak thesis and a strong thesis – the former stating "I am able to do something such that, if I did it, a law would be broken" and the latter stating "I am able to break a law" (Lewis 1981, p.115) – and claims a compatibilist, whilst rejecting the strong thesis, could readily accept the weak one. When discussing an example of the possibility of having raised his hand when he did not in fact do so, Lewis has the following to say:

[84] There are, of course, a number of other arguments against the conclusions of the Consequences Argument. One such focusses on a "hypothetical" analysis of the words "can" or "ability" which interprets "Jessica is able to stop and help" as meaning "Jessica *would* have stopped to help if she had *wanted* to." Note, this can be true even if what she *wanted* was determined (see Kane 2005, p.26-28 for a general introductory discussion on this). However, I take Lewis' argument to be the more persuasive.

> Now consider the disputed case. I am able to raise my hand, although it is predetermined that I will not. If I raised my hand, some law would be broken. I even grant that a law-breaking event would take place. (Here I use the present tense neutrally. I mean to imply nothing about *when* a law-breaking event would take place.) But is it so that my act of raising my hand would cause any lawbreaking event? Is it *so* that my act of raising my hand would itself be a law-breaking event? Is it so that any other act of mine would cause or would be a law-breaking event? If not, then my ability to raise my hand confers no marvelous ability to break a law, even though a law would be broken if I did it. (Lewis 1981, p.116)

Had what Lewis refers to as a “divergence miracle” (ibid., 117) occurred prior to him raising his hand – which resulted in a miraculous diverging from the state of affairs where he could not raise his hand to a state of affairs where he could – a law would indeed have been broken, but this law breaking would not have been *caused by him raising his hand*. The occurrence of such a “miracle“ does not mean there has been any violation of a law in any given world, just that the world in question could deviate from what would usually happen in accordance with the laws of worlds which have the same history as that world. Thus, it is a miracle relative to those worlds, not relative to its own. Lewis’ point is that the implications for the agent that van Inwagen and the other supporters of the Consequence Argument are claiming do not follow. The incompatibilists (including the libertarians) would wish to argue that compatibilists are committed to the claim that for Jessica to have the ability to stop and help – when what she actually did was drive on – requires her to *cause* that the laws of nature (or the past) be different, when really compatibilists are committed only to the claim that for Jessica to have the ability to stop and help – when what she actually did was drive on – requires only that the laws of nature (or the past) *be different*; and, further, had she acted otherwise, one or other *would have been different*. There is no requirement for this breaking of the laws to be caused by the agent, thus they require no special power in order to be able to do so, and the Consequence Argument fails.

This, of course, remains an open debate (and van Inwagen and others have made attempts to rebut Lewis’ objection) but what this brief discussion shows is that there is a clear sense of how things could have been otherwise which in no way assumes indeterminism and thus, contrary to what the libertarians believe, it is by no means indisputable that indeterminism is *necessarily* required to meet the AP condition.

Ultimate Responsibility

But what of the other problematic condition – Ultimate Responsibility (UR) – which has also caused libertarians to believe it is essential that they incorporate indeterminism into their models? This condition effectively states that if an agent is to be ultimately responsible for any act, she must also be responsible for any preceding actions, decisions or events which can be said to be causally responsible for that act. By extension – as both libertarians, such as Kane, and free will sceptics, such as Strawson, argue – if an agent is to be considered fully responsible for acts determined by her *will* (or current established psychology, CEP), she must be responsible for the formation of the will from which such acts flow: she must be (at least in some sense) *causa sui*. However, as we have seen in our consideration of the Basic Argument, this then seems inevitably to result in a problematic regress; with the CEP that determines an agent's actions needing to be intentionally and voluntarily formed by actions determined by a previous established psychology (EP), which then needs to be formed (voluntarily and intentionally) by actions determined by a further previous EP, and so on and so forth. We can never actually, therefore, truly be the cause of ourselves – can never be *ultimately* responsible for the formation of our character which determines how we will act – at any stage.

And this argument applies whatever the fundamental nature of the universe. If the universe were to turn out to be deterministic then *every* event would have a sufficient cause and the regress could be traced back indefinitely into the past, reducing to an unacceptably low level (in the view of the libertarians at least) any potential for assigning applicable responsibility to the eventual adult. If the universe were to turn out to be indeterministic an inevitable regress of decisions made by a CEP, which is formed from earlier decisions made by a previous EP still remains, and any attempts to explicitly utilise indeterministic factors, via the introduction of deliberate causal breakages – where undetermined voluntary actions can play a part in such a process of will-formation – seemingly just introduces a significant amount of arbitrariness which, again, reduces to an unacceptably low level any potential for assigning applicable responsibility to the eventual adult.

Robert Kane, who proposes his self-forming actions (SFAs) as contenders for such voluntary, will-setting actions (see section 2.3), argues that UR is actually the more important condition for libertarians – more so than the more historically discussed AP – and therefore meeting it satisfactorily is essential in order for an agent to achieve the supposed Gold Standard freedom of the will. Furthermore, it is his contention that UR *entails* both indeterminism and AP, as without indeterminism all our actions would have sufficient causes and without AP there could not be any will-setting actions as these require the agent to be able to do otherwise (in Kane's "plural voluntary" sense – discussed in section 2.3 – which is to do otherwise intentionally and voluntarily and not inadvertently). For Jessica to be ultimately responsible for her decision to, say, abandon the fallen man to his fate, (assuming such a torn decision would be of the correct sort to instigate an SFA) she must not be sufficiently caused to act in this way by antecedent conditions, thus her decision must be undetermined (requiring indeterminism) and she must have genuine alternatives available which she can take advantage of in a way that satisfies Kane's plurality conditions, which requires AP. But in the same way that the AP condition drives the Luck Objection, the UR condition drives the Basic Argument, and (as we have seen throughout this thesis) any attempt to counter one seems simply to strengthen the grip of the other. Libertarians do want our decisions and actions to be broadly determined by our wills – because this is what gives the agent sufficient control for an assigning of responsibility – but this leads to a regress (and the Basic Argument) so they introduce causal breakages which (at least in Kane's theory) allow for voluntary, undetermined, will-setting actions, but this leads to accusations of arbitrariness (and the Luck Objection); and so on, back and forth – hence the *Catch-22 of Libertarianism*. But is indeterminism (which I have already shown is arguably not required to meet the AP condition) actually *essential* to achieve UR, or (at least) a satisfactory version of it?

Let us now return to our earlier discussion where we considered our willingness (or lack of such) to allocate any blame to the five-year-old Jessica – just starting out on her athletics career at the direction of her father – for any actions that could potentially play a causal role in the formation of an adult character (which may one day choose to leave an elderly man collapsed by the side of the road). In doing so, we can start to consider this libertarian need to incorporate indeterminism in order to meet the UR condition more closely. The first thing to notice about such situations is that we seem intuitively to hold an adult responsible for decisions made from a character we assume is formed (at least in part) from decisions they make as a child, whilst *not* holding the child responsible for *their* formative decisions in the same way – even though some of these decisions we assume can be made fully autonomously, especially as they get older. Our instincts when considering issues of development and responsibility, therefore, may not in fact tally with this notion of some kind of full (causal) regression from pseudo-responsible adult, back through necessitating formative influences to a morally incapable child, that arguments such as the Basic Argument are trading on, and we do in fact naturally attribute more self-determination in the formative process.

This is an argument made by Joel Feinberg in his essay “The Child’s Right to an Open Future”, in which, after a wider discussion of the influence of parents and the state in the rearing of a child and the importance of not overly limiting their developmental opportunities and possibilities to travel down a variety of roads in life, Feinberg addresses the potential paradoxes of self-fulfilment and self-determination that his discussion raises. He argues that paradoxes such as the seemingly plausible picture of a necessarily infinite series of prior selves that appears to result from any attempt to attribute genuine self-determination to an agent is really just based on “approximate generalisations” and “partial truths” (Feinberg 1992, p.95).

> It is an overstatement, for example, that there is any early stage at which a child’s character is *wholly* unformed and his talents and temperament *entirely* plastic, without latent bias or limit, and another that there can be *no* “self-determination” unless the self that does the determining is already fully *formed*. Moreover, it is a distortion to represent the distinction between child and adult in the rigid manner presupposed by the “paradoxes”. (Ibid., p.95)

The child Jessica (even at five years old) is not entirely a passive receptor of her father’s influences, which would then be considered as sufficient causes for some early self in an unbroken chain of succeeding selves. Jessica herself has some level of unadulterated input into her own development from birth as genetic proclivities play their role in the constant evaluation and assimilation of external influences that develops her CEP and there is no fixed point at which she suddenly becomes a fully-formed adult self, solely responsible for all future character development. It is in acknowledging these facts of human development that Feinberg believes the paradoxes – such as that of an inevitable vicious regress – can be tempered.

> Thus from the beginning the child must – inevitably *will* – have some “input” in its own shaping, the extent of which will grow continuously even as the child’s character itself does. I think that we can avoid, or at least weaken, the paradoxes if we remember that the child can contribute toward the making of his own self and circumstances in ever-increasing degree. Always the self that contributes to the making of the new self is itself the product of both outside influences and an earlier self that was not quite fully formed. That earlier self, in turn, was the product of both outside influences and a

> still earlier self that was still less fully formed and fixed, and so on, all the way back to infancy. At every subsequent stage the immature child plays an ever-greater role in the creation of his own life, until at the arbitrarily fixed point of full maturity or adulthood, he is at last fully and properly in charge of himself, sovereign within his terrain, his more of less finished character the product of a complicated interaction of external influences and ever-increasing contributions from his own earlier self. At least that is how growth proceeds when parents and other authorities rear a child with maximal regard for the autonomy of the adult he will one day be. That is the most sense that we can make of the ideal of the "self-made person," but it is an intelligible idea, I think, with no paradox in it. (Ibid., p.97)

It can, of course, be argued (and perhaps would be so by Kane and/or Strawson) that the initial conditions of heredity and environment will (Strawson), or cannot be allowed to (Kane), fully dictate every stage of future self through which the child will pass on the way to adulthood. But what Feinberg is offering is a different picture of development which is independent of the fundamental physics of the universe, and in which a child does have a meaningful and efficacious role in the development of their own character. The initial genetic conditions, rather than being viewed as the first step on a fully determined road, provide nothing more than a biased starting point that necessarily functions as the child's first input into their own development. From then on, the child's development naturally assimilates all sorts of influences and experiences which filter through and affect her ever-developing character, or CEP, which can be seen as essentially in control of *how* such influences do affect her character formation.

The picture the Basic Argument is presenting is one of a fully formed psychological character (CEP) at every stage making a decision, which then produces the next fully formed CEP and which can thus be traced backwards – by way of a vicious regress – to a point of no assignable agential responsibility, and the explicit focus on indeterminism by libertarians such as Kane as a solution to this problem then raises the importance of the debate surrounding the fundamental framework of the universe. The alternate picture Feinberg is presenting is one of initial biases which would not necessarily determine anything, but which allow for constant processes of active influence/experience evaluation and cognitive affectation that allows for self-development in a non-contradictory manner that is independent of whether it should turn out that the true nature of physics is deterministic or indeterministic. Jessica is not born as either some kind of *tabula rasa* or a fully determined potential adult waiting to develop, but as an active player with plenty of preinstalled dispositions and capabilities that *could* lead to her quite possibly refusing at age five to play any part in her father's athletic intentions for her. And the ever-evolving CEP formulating her choices to join athletics clubs, train, compete and then perhaps leave an elderly man collapsed on the floor, can indeed be seen as much *her* creation as it is by anything else.

Similarly, therefore, to the way Lewis distinguishes between the weak thesis and the strong thesis for the AP, or "could have done otherwise," principle, we can distinguish between a weak and strong version of UR.

> **The Strong Version** - If an agent is to be ultimately responsible for any act, she must also be fully responsible for any preceding acts, decision or events which can be said to have sufficiently caused that act.
>
> **The Weak Version** – If an agent is to be considered fully responsible for acts determined by her *will* (or CEP), she must be responsible for the formation of the will from which such acts flow.

These two versions are most often presented as equivalent in the literature and, as such, it is argued (in the same way it is argued *contra* Lewis) that any person committed to one must be committed to the other. However, one makes a statement about causal sufficiency and one does not; thus, as the above discussion of Feinberg's work has just shown, the weak version can be met irrespective of whether the universe is deterministic or indeterministic. In the same way that the weak thesis of the AP condition can be met whatever the fundamental nature of the universe, we can also meet the weak version of UR and, in this way, an agent *can* have alternate possibilities *and* be ultimately responsible for their decisions and actions however the physics should turn out to be. But if indeterminism is not required to meet either the AP or UR conditions, need it feature at all in any account of freedom and responsibility?

4.3 A New Gold Standard and a New Way Forward

In trying to use indeterminism as a means to meet both the AP and UR conditions (and thus to meet principles 3, 4 and 5) the philosophers and scientists whose work I have considered in this thesis have, I believe, fallen into an unresolvable contradictory position, being led down paths that end largely in confusion, incoherence and/or mystery. Yet, as I have just argued, indeterminism is not necessarily required to meet *either* of these two key conditions. Furthermore, I believe that what the discussions throughout this thesis have demonstrated (somewhat to the contrary to how it is generally portrayed) is that the indeterminism featured so heavily in the included libertarian models is not actually doing anything conceptually important. It is not, in fact, *doing any positive work at all* and thus all the well-established confusions its inclusion creates could arguably be said to be entirely unnecessary. In fact, I contend that if we were to remove all references to indeterminism completely from all of the featured theories (as well as from any discussion of the associated implications) they would all – as substantive models for decision-making and action – remain completely unaffected. Instead of being viewed as some essential and active ingredient in libertarian theories, it should arguably be seen as irrelevant to the accounts of freedom and responsibility such theories produce.

4.3.1 The Proposed Models Reimagined

To make good on my contention, let us consider Robert Kane's event-causal approach (which I take to be one of the best attempts to provide a coherent libertarian account of free will). Kane's model gains its intuitive appeal based as it is on the causation of physical events (such as those connected with thoughts and actions) by other associated physical events. This event-causal foundation is, of course, precisely what compatibilist theories also rest on and those theories, it could be argued, present their version of event-causation in a more *complete* fashion – without any apparent need to posit any causal breakages. It is for Kane, then, to show what his theory (and its prominent incorporation of indeterminism) offers on top of the compatibilist picture that would make it preferable. In truth, however, it offers *nothing* extra because a close reading of his work shows that whether the universe turns out to be deterministic or indeterministic is entirely irrelevant to Kane's account of deliberation and decision-making. If you were to remove any reference to the presence of indeterminism from Kane's model the theory is wholly unaffected and this is because, other than simply stipulating that the universe is required to be indeterministic and that certain outcomes are thus indeterminate (in order to comply with the key libertarian conditions), Kane's approach is just a well-developed, well-presented, and instinctively attractive model for decision-making which anyone could endorse, *including compatibilists*.

Kane's approach to decision-making (we can remind ourselves) breaks down as follows:

1) Information regarding a particular dilemma is received into a central processing area (mediated by the agent's CEP), which requires evaluation, the formulation of a decision and, ultimately, action.
2) The CEP decides, upon evaluation, whether the dilemma under consideration is straight-forward enough for the decision to be generated mechanically from the CEP.
3) If it is not (perhaps due to a moral ambiguity or novelty of the situation), then the self-forming action process activates within the self-network – the consequence of which is a division within the agent's mental processing function that produces the two "possibility streams" (parallel processing) that vie for seniority, creating genuine quantum indeterminism along the way.
4) One of these streams then passes some threshold and becomes *the* decision, and the relevant action is brought about.
5) The agent then acknowledges the action as *her* action and sees it as the implementation of something she was *trying to achieve.*

The main thing to note with this structure is that the featured indeterminism, as already stated, is not adding anything substantive to the model. Indeed, if you remove the words "creating genuine quantum indeterminism along the way" the process remains exactly the same; and thus the indeterminism in Kane's account appears to be an unnecessary addition bolted on solely to comply with the basic libertarian requirements for achieving a "true" freedom of the will.

Consider the following amended version:

1) Information regarding a particular dilemma is received into a central processing area (mediated by the agent's CEP), which requires evaluation, the formulation of a decision and, ultimately, action.
2) The CEP decides, upon evaluation, whether the dilemma under consideration is straight-forward enough for the decision to be generated mechanically from the CEP.
3) If it is not (perhaps due to a moral ambiguity or novelty of the situation), then the self-forming action process activates within the self-network – the consequence of which is a division within the agent's mental processing function that produces the two "possibility streams" (parallel processing) that vie for seniority.

4) One of these streams then passes some threshold and becomes *the* decision, and the relevant action is brought about.
5) The agent then acknowledges the action as *her* action and sees it as the implementation of something she was *trying to achieve.*

It is clear that this amended version, which has simply undergone the specified minor alteration – and which remains an intuitive sketch of human decision-making – is equally compatible with either determinism *or* indeterminism. It makes no difference which should turn out to be true. The reason, therefore, that Kane's approach is often considered to be broadly intuitive and straightforward (in a similar way to how many view the compatibilist approach) is simply because it *is* intuitive and straightforward, but only so long as you leave the extraneous indeterminism out of the model. Shoe-horning indeterminism explicitly into the picture as some kind of prominent force (needed to free agents from the assumed invisible shackles that come with the deterministic notion of a single, causally necessitated future) results in the confusions and incoherence; all the while adding nothing in return.

Other proposed models which I have evaluated in this thesis can be looked at in a similar fashion. Consider the schematic Timothy O'Connor proposes when discussing how his agent-causal model can account for an agent's actions in terms of her reasons for acting:

> The agent acted then in order to satisfy his antecedent desire that Θ if
>
> 1. prior to this action, the agent had a desire that Θ and believed that by so acting he would satisfy (or contribute to satisfying) that desire;
> 2. the agent's action was initiated (in part) by his own self-determining causal activity, the event component of which is the-coming-to-be-of-an-action-triggering-intention-to-so-act-here-and-now-to-satisfy-Θ;
> 3. concurrent with this action, he continued to desire that Θ and intended of this action that it satisfy (or contribute to satisfying) that desire; and
> 4. the concurrent intention was a direct causal consequence (intuitively, a continuation) of the action-triggering intention brought about by the agent, and it causally sustained the completion of the action. (O'Connor 2000, p.86)

For this description of O'Connor's model, no changes at all need to be made for it to be compatible with determinism. What we have, once again, is just an intuitive model for human decision-making and action based on desires held. There is no requirement for O'Connor to include in this schematic the added elements that the desire cannot cause the action or that the agent herself cannot be caused to do anything, by anything, because such extra commitments do not add anything relevant to the model. Such a schematic works perfectly well in either a deterministic or indeterministic setting.

Carl Ginet's non-causal model is essentially the same as O'Connor's (minus the agent-causal capacity) so his would also work just as well in either a deterministic or indeterministic universe. Peter Tse, himself, admits that his Criterial Causation model would work just as well within a deterministic framework and it can easily be argued that the fundamental nature of the neuronal inputs arriving at the synapses (the only place where an indeterministic process is allowed/required on Tse's account) is largely irrelevant to the overall picture of human agency Tse has developed.[85] Roger Penrose (as discussed above in section 3.3) is more interested in the non-computability of his consciousness-generating wavefunction collapse which does not necessarily require indeterminism and thus the Orch OR model, as applied to the topic of free will, would (yet again) work just as well in either setting. Though Henry Stapp is clearer when it comes to believing the universe is fundamentally indeterministic in nature, it also remains true, however, that Processes 0 through to 3 could all still function the same where it to transpire that he is incorrect.

Thus, not only is indeterminism *not* required to meet the AP and UR conditions, its removal from all the libertarian models leaves them completely intact.

4.3.2 The *Criteria* (One Final Assessment)

A further question now presents itself: would an acceptance of the irrelevance of indeterminism mean that the Gold Standard is to be forever out of reach? I do not believe so and, to demonstrate this, let us return, for one final time, to my *Criteria for Coherence* which (it will be recalled) was initially derived from the key tenets of what libertarians believe is required to achieve the Gold Standard.

[85] See Tse 2013, p.145 §7.18

P1 – The theory must be a true libertarian (incompatibilist) theory: it must proceed on the assumption that the desired Gold Standard form of free will *exists* and is incompatible with determinism.
P2 – The theory must therefore rest on a platform of indeterminism: it cannot be that (all) our decisions and actions are wholly necessitated by antecedent states or events.
P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.
P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.
P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.
P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

The first two principles are based on the libertarian belief that indeterminism *must* be included in any model for free action in order to achieve the Gold Standard freedom of the will. If we remove those entirely, we are left with the following revised criteria:

P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.
P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.
P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.
P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

Now, let us bring back in Kane's model for action/decision-making, as modified by me to remove any reference to indeterminism.

1) Information regarding a particular dilemma is received into a central processing area (mediated by the agent's CEP), which requires evaluation, the formulation of a decision and, ultimately, action.
2) The CEP decides, upon evaluation, whether the dilemma under consideration is straight-forward enough for the decision to be generated mechanically from the CEP.
3) If it is not (perhaps due to a moral ambiguity or novelty of the situation), then the self-forming Action process activates within the self-network – the consequence of which is a division within the agent's mental processing function that produces the two "possibility streams" (parallel processing) that vie for seniority.
4) One of these streams then passes some threshold and becomes *the* decision, and the relevant action is brought about.
5) The agent then acknowledges the action as *her* action and sees it as the implementation of something she was *trying to achieve.*

Using this model, and the concepts we have been discussing concerning the development of a child into adulthood, we can build a fuller picture of our athlete example.

Jessica, our budding athlete, is born with nothing more than genetic predispositions to (amongst many other things) heightened athletic prowess and training application. Throughout her formative years, she is encouraged to train and compete but, at every stage, is involved in and at least partly in control of her own development (to an ever-greater degree). As such, she assimilates and evaluates the influences around her – along with her own experiences – within her ever-evolving CEP and makes decisions and acts accordingly. Her character continues to develop in this way under the guidance of her CEP with some decisions taken instinctively, some requiring more thought, and some requiring prolonged and extensive deliberations, until that fateful day when she is on her way (late) to the Olympic qualification race. As she sees the elderly man fall to the pavement the associated sensory information is processed by her CEP and flagged as requiring an important decision to be made, the nature of which is sufficiently ambiguous as to require an SFA. Jessica's mental processes divide themselves into two parallel possibility streams – one of which would result in her helping, the other of which would result in her driving on to her race – that vie with each other (presumably against a backdrop of her desires, needs, beliefs and so forth) for supremacy. The stream resulting in her helping the man wins out and Jessica exits her vehicle and runs over to help. When asked later at the hospital by the man's relatives what made her give up her race and come to his aid, she answers: "I just knew it was the right thing to do."

The question to put back to proponents of the Luck Objection and the Basic argument now is: in what way does Jessica's development (as outlined in this scenario) not meet the standard we have set regarding the ability to *choose* our own lives – the standard which meets both the AP and UR conditions and focusses on the agent's own opportunities rather than the perceived implications of the movements of subatomic particles? Those who see the UR condition as particularly problematic would no doubt press the point that I have acknowledged the importance of genetic predispositions and no one can "choose" their genetic makeup, which they say is what ultimately sets us out on our inevitable causal path (if not on our random adventure). Firstly, it could be argued that this is a mistaken view of our "genetic self" as surely we just *are* our genetics (however they come about) and thus the "buck" does arguably stop there but, secondly (and more importantly), this notion that we do not choose our genetics misses the real factors that we *actually use* to judge responsibility – which is much more the Feinberg*ian* notion of available and affected opportunities – and was never part of the standards employed to judge moral development. Insisting that the mere existence of genetic predispositions instigate development chains that remove responsibility from any adult agent is to invoke the strong version of UR and claim its primacy, but for what reason should it be accepted as more valid than the weak version?

We can now evaluate this scenario against the revised criteria.

P3 – At any point a libertarian agent must be free to act or act otherwise, whatever the past conditions and natural laws.

Jessica, as we have seen from our discussion of the AP condition (and Lewis' interpretation of it), is fully able to meet the AP condition – and thus is able to act or act otherwise – regardless of whether the universe should prove deterministic or indeterministic. Thus, principle 3 is met.

P4 – The theory must provide a model for decision-making and action which is neither fully determined *nor* arbitrary, thus affording the agent full control over their acts in such a way as to ensure responsibility can be justifiably ascribed.

Jessica makes a decision based on her CEP as it is at that time. As our discussion of the UR condition (and Feinberg's analysis) showed, at no time is the formation of her CEP fully determined by factors outside of her control and, as her CEP is formulating the ultimate decision, nor is this decision (and thus her previous decisions and character formation in general) fully subject to arbitrary factors. Her actions are thus within her control and she can be held responsible for them. Jessica has a character from which her decision results and neither the character formation nor the resultant decision need rely on either determinism or indeterminism in order to confer satisfactory control. Principle 4 is therefore also met.

P5 – The theory must explain how an agent can be the cause of herself without leading to an infinite regress, thus ensuring the locus of control lies fully within *her* (within her CEP) and not within something else external to her.

As above, Jessica can indeed be the cause of herself in a perfectly satisfactory sense without leading to an infinite regress. The weaker version of UR is all anyone need be committed to and this just states that an agent need be responsible for the formation of her CEP, character, or *will*, from which her actions flow. Right through from being born with *some* genetic predispositions, to assimilating influences and experiences and making decisions from early on (albeit perhaps *guided* initially), on to a more autonomous adulthood, Jessica is continually going through the process of self-formation. But this addition of external/environmental/influential factors removes the inevitability of any problematic regress and the fact that these external factors are constantly reviewed and assimilated by *her* – via her CEP – means she retains the locus of control within her. Principle 5 is also met.

P6 – The theory must avoid any invocation of overly obscure or inscrutable forms of agency or causation or over-reliance on mysterious interventions.

As there is no need to explicitly incorporate indeterministic process in order to meet the criteria above, there is no need to invoke any obscure or mysterious forms of causation or of other neurological processes. I contend there is therefore nothing incomprehensible about the model described above and it contains no unusual ontological implications. Principle 6 is met.[86]

This revised model – which (unlike Kane's) does not require any explicit inclusion of indeterministic processes – demonstrates that we can in fact retain all the positive features of the best libertarian theories *whatever the fundamental physical nature of the universe should turn out to be*. As such, it is able to meet all the principles of the revised *Criteria* (which still represents everything substantively important that the libertarians state they wish to achieve) and thus meet all the important goals of the libertarian project in an entirely non-contradictory and coherent fashion. This is something the libertarian models themselves are not able to do specifically *because* their belief that their own criteria *can only* be met within the framework of an indeterministic universe forces them to employ particular strategies to meet them that result in such contradictions. I have argued that this libertarian belief that indeterminism is categorically *required* to meet their stated goals is simply mistaken but, to be clear, this does not mean I am just arguing for compatibilism (at least not in its traditional form). It could very well be suggested that – in contrast to the "impossibilists" (such as Galen Strawson) who believe that any genuine freedom of the will must be incompatible with both determinism *and* indeterminism – my position is a "possibilist" approach, which argues that a genuine freedom of the will should be considered to be compatible with *both* potential physical frameworks. But what I am *really* arguing is that whether the universe turns out to be deterministic or indeterministic should, in fact, just be considered as wholly *irrelevant* to the issue of freedom and responsibility – so perhaps we should label me an "irrelevantist."

Whichever label we ultimately deem to fit best, the point is that questions concerning fundamental physics simply need not be considered as part of the debate on free will, and this position is supported (at least in part) by my above demonstration that all the models for free action contained within all the libertarian accounts I have considered are able to keep all their relevant components – entirely coherently – without any commitment to indeterminism. Indeed, as we have seen that no explicitly

[86] It could perhaps be argued that my position of effectively ignoring the problems posed by indeterminism and determinism – of arguing that agency and action is somehow divorced from the fundamental causal order – is itself adding a layer of obscurity, but I would dispute that this is in fact the case. The position only becomes obscure if you try and relate it to that fundamental level, which is precisely what I am arguing against doing. I am in fact deliberately arguing that we do not need to involve any strong metaphysics such as determinism or indeterminism at all as all the positives in the libertarian accounts do not trade on whether determinism is or is not true.

indeterminist theory can be made to be coherent, and also that the addition of indeterminism adds nothing substantive to the models the libertarians have devised (plus the fact that my revised model meets all the key requirements libertarians say they have), for those left unconvinced by the traditional compatibilism of committed determinists or unwilling to take the hard line of hard determinism or impossibilism, irrelevantism may be the only avenue left open to them.

4.3.3 Physics, Agency and a New Way Forward

The concepts of determinism and indeterminism – and the implications of adopting either as a fundamental feature of our universe – have loomed so large over debates concerning free will and responsibility for so great a period of time that any suggestion they should be considered irrelevant to the debate may be a difficult pill to swallow. In the long history of the debate, theorists have (more often than not) fallen into one camp or another and then worked on trying to reconcile intuitive notions concerning freedom and responsibility with their favoured fundamental physical framework. This approach to the issue has often led to the debate being reduced to one pitted against the other, with the higher-level models for the actual workings of human agency and action coming secondary to this need to overcome the apparent problems generated by whichever 'foundational' principles are adopted. The philosophers and scientists whose work I have considered in this thesis have seemingly followed this same road and, in trying to explicitly incorporate indeterminism in particular as a fundamental and prominent component of their models, have (as I have argued) got themselves into an unresolvable contradictory position.

But, given the forgoing discussion, perhaps it really is time the free will debate moved on from this encumbering framework of "Determinism vs Indeterminism" or – more specifically – from the argument over which of the two is *more* compatible with free action and gave serious consideration to this notion that they should both be considered as largely irrelevant to discussions of freedom and responsibility, with neither bearing any relevance to substantive models of agential action and decision-making. The common tendency to view determinism and indeterminism as some kinds of universal forces (often malign) that have the power to *interfere* in the lives of human beings is, I would argue, simply mistaken. This does not mean, however, that I am attempting just to close the door on the monster in my house and dealing with it by pretending it simply is not there.[87] Rather, I am saying the monster was never *really* there to begin with. If anything, it is nothing more than a harmless stuffed toy, minding its own business in the corner of the room, which only grew into a fanged and terrifying beast capable of destroying human freedom in the imaginations of philosophers throughout the ages. Indeed, the historical fetishisation of these scientific notions – which are really nothing more than potential physical frameworks for the motions of particles or waves and can thus be relegated to issues concerning the prediction and calculations within physics – has arguably hindered progress in the development of a more appropriate conceptual framework which could be applied in order to further the debate on freedom and responsibility; one that focuses more on *agency* as a psychological phenomenon, and one which has its own terms to describe mental events that should not necessarily be seen as automatically, ontologically inferior to those proffered by the physical sciences.

As Crane and Mellor point out when arguing that the exclusion of such psychological concepts leaves physicalism as an inevitably “vacuous doctrine about the mind”:

> Something about the mental is supposed to deprive psychology of the ontological authority of physics and chemistry. But what? What prevents psychology from telling us in its own terms what kinds of mental things and events there are? (Crane and Mellor 1990, p.187)

Physics, it could well be argued, is simply not in the business of saying anything about agency, which does not figure as a notion in any laws of nature, and, by return, philosophers debating free will have no need to concern themselves with fundamental physics. As such, Daniel Dennett’s intendedly derisory phrase “moral levitation” – used to attack the libertarian position – is (in a sense) entirely correct as a description of agency; but only in regards to its conceptual distinction from any potential foundational science of human action. And, as Crane and Mellor discuss, why should the physical sciences be given priority (or more “ontological authority”) over such disciplines as psychology when it comes to the nature of reality in general and the notion of agency in particular?

Libertarians, therefore, might consider the switch to irrelevantism and in doing so could move on from complicated and contradictory attempts to include the “deep openness” of an absolute indeterminism explicitly in their models of free action, safe in the knowledge that they would *not* just be left with a

[87] My thanks to Sam Coleman for the analogy.

problematic deterministic picture and all the perceived negative connotations (metaphysical limitations and constrictions) that such a framework is often seen to bring with it. Rather, they can just discard such concerns entirely, retain all the positives of their proposed models, and work instead on expanding the explanatory network of the concept of agency; thus putting the more *human* sciences on a par with the physical ones.

In conclusion: Is libertarian free will an inescapably incoherent concept? Yes, but only because it insists on the explicit inclusion of indeterminism. Removal of all traces of this scientific notion leaves each of the models for free action I have considered with a real chance of being considered fully intact, fully coherent, and fully retaining of all the important qualities libertarians hold dear. Libertarian free will **is** ***incoherent. Free will is not necessarily so.***

References

Austin, J. L. (1956). “Ifs and cans.” In *Proceedings of the British Academy, vol. 42*. pp. 109-132.

Chappell, Vere (ed.) (1999). *Hobbes and Bramhall on Liberty and Necessity*. Cambridge University Press.

Chisholm, Roderick M. (1964). “Human Freedom and the Self.” Reprinted in *Free Will*, ed. Gary Watson, 26-37. New York: Oxford University Press, 2003.

Chisholm, Roderick M. (1967). “He could have done otherwise.” *Journal of Philosophy* 64(13): 409-17.

Clarke, Randolph. (2003). *Libertarian Accounts of Free Will*. New York: Oxford University Press.

Clarke, Randolph. (2010). “Because She Wanted To.” *The Journal of Ethics,* 14(1): 27-35.

Clarke, Randolph. (2010a). “Agent Causation.” In *A companion to the philosophy of action*, eds. Sandis, Constantine, and Timothy O’Connor, 218-226. Oxford: Blackwell.

Clarke, Randolph. (2011). “Alternatives for Libertarians.” In *The Oxford Handbook of Free Will*, ed. Robert Kane, 329-348. New York: Oxford University Press.

Crane, Tim & Mellor, D. H. (1990). There is No Question of Physicalism. *Mind* 99 (394):185-206.

Dennett, Daniel C. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. London, England: MIT Press.

Dennett, Daniel C. (2003). *Freedom evolves*. New York: Viking Press.

Feinberg, Joel (1992). *Freedom and Fulfillment: Philosophical Essays*. Princeton University Press.

Fischer, John Martin. (2001). “Review of *Persons and Causes* by Timothy O’Connor.” *Mind* 110(438): 526-531.

Fischer, John Martin. (2014). “Toward a Solution to the Luck Problem.” In *Libertarian Free Will: Contemporary Debates*, ed. David Palmer, 52-68. New York: Oxford University Press.

Ginet, Carl. (1990). *On action*. Cambridge: Cambridge University Press.

Ginet, Carl. (1997). Freedom, responsibility, and agency. *The Journal of Ethics* 1 (1):85-98.

Ginet, Carl. (2008). “In Defense of a Non-Causal Account of Reasons Explanations.” *The Journal of Ethics* 10: 229-237.

Ginet, Carl. (2016). “Reasons Explanation: Further Defense of a Non-causal Account.” *The Journal of Ethics,* 20: 219-228.

Grush, Rick & Churchland, Patricia Smith. (1995). “Gaps in Penrose's toilings.” *Journal of Consciousness Studies* 2 (1):10-29.

Hameroff, Stuart. (2012). "How quantum brain biology can rescue conscious free will". *Frontiers in Integrative Neuroscience*. 6: 93.

Hameroff, Stuart & Penrose, Roger. (1995). What 'gaps'? Reply to Grush and Churchland. *Journal of Consciousness Studies* 2 (2):98-111.

Hameroff, Stuart & Penrose, Roger. (1996). Orchestrated objective reduction of quantum coherence in brain microtubules: The "Orch OR" model for consciousness. *Mathematics and Computers in Simulation* 40:453-480.

Hameroff, Stuart & Penrose, Roger. (2013). “Consciousness in the universe. A review of the ‘Orch OR’ theory.” *Physics of Life Reviews* 11: 39-78.

Hiddleston, Eric. (2005). “Review of *Persons and Causes* by Timothy O’Connor.” *Nous* 39(3): 541-556.

Honderich, Ted. (1993). *How Free Are You?* Oxford: Oxford University Press.

Kane, Robert. (1996). *The Significance of Free Will*. New York, US: Oxford University Press.

Kane, Robert. (1999). “Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism.” *The Journal of Philosophy* 96: 217-240.

Kane, Robert. (2002). “Some Neglected Pathways In the Free Will Labyrinth.” In *The Oxford Handbook of Free Will*, ed. Robert Kane, 406-437. New York: Oxford University Press.

Kane, Robert. (2005). *A Contemporary Introduction to Free Will*. Oxford: Oxford University Press.

Kane, Robert. (2011). “Rethinking Free Will: New Perspectives On An Ancient Problem.” In *The Oxford Handbook of Free Will*, ed. Robert Kane, 381-404. New York: Oxford University Press.

Kane, Robert, (ed.) (2011). *The Oxford Handbook of Free Will*. New York: Oxford University Press.

Kane, Robert. (2014). “New Arguments in Debates on Libertarian Free Will: Responses to Contributors.” In *Libertarian Free Will: Contemporary Debates*, ed. David Palmer, 179-214. New York: Oxford University Press.

Kane, Robert. (2019). “The complex tapestry of free will: striving will, indeterminism and volitional streams.” *Synthese* 196: 145-160.

Levy, Neil. (2013). “Peter Ulric Tse, *The Neural Basis of Free Will: Criterial Causation*. Review.” *Philosophy in Review* 33(4): 331-333.

Lewis, David. (1981). “Are we free to break the laws?” *Theoria* 47 (3):113-121.

Mele, Alfred R. (1992). *Springs of action: understanding intentional behavior*. New York: Oxford University Press.

Mele, Alfred R. (1998). “Review of Robert Kane's *The Significance of Free Will*.” *Journal of Philosophy* 95 (11):581-584.

Mele, Alfred R. (2006). *Free Will and Luck*. New York, US: Oxford University Press.

Mele, Alfred R. (2014). "Kane, Luck, and Control: Trying to Get by without Too Much Effort." In *Libertarian Free Will: Contemporary Debates*, ed. David Palmer, 37-51. New York: Oxford University Press.

O'Connor, Timothy. (1993). "Indeterminism and free agency: Three recent views." *Philosophy and Phenomenological Research* 53(3): 499-526.

O'Connor, Timothy, ed. (1995). *Agents, causes, and events: Essays on indeterminism and free will*. New York: Oxford University Press.

O'Connor, Timothy. (1995b). "Agent Causation." In *Agents, causes, and events: Essays on indeterminism and free will,* ed. Timothy O'Connor, 173-200. New York: Oxford University Press.

O'Connor, Timothy. (2000). "Causality, Mind and Free Will." *Philosophical Perspectives* 14: 105-117.

O'Connor, Timothy. (2007). "Is it all just a matter of luck?" *Philosophical Explorations* 10(2): 157-161.

O'Connor, Timothy. (2009a). "Degrees of freedom." *Philosophical Explorations* 12(2): 119-25.

O'Connor, Timothy. (2010). "Reasons and Causes." In *A companion to the philosophy of action*, eds. Sandis, Constantine, and Timothy O'Connor, 129-138. Oxford: Blackwell.

O'Connor, Timothy. (2011). "Agent-Causal Theories of Freedom." In *The Oxford Handbook of Free Will*, ed. Robert Kane, 309-328. New York: Oxford University Press.

Penrose, Roger. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.

Penrose, Roger. (1999). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press.

Stapp, Henry P. (1993). *Mind, Matter and Quantum Mechanics*. New York: Springer-Verlag.

Stapp, Henry P. (1999). "Attention, intention, and will in quantum physics." *Journal of Consciousness Studies* 6(8-9): 8-9.

Stapp, Henry P. (2001). "Quantum Theory and the Role of Mind in Nature." *Foundations of Physics* 31(10): 1465-1499.

Stapp, Henry P. (2005). "Quantum Interactive Dualism: An Alternative to Materialism." *Zygon* 41 (3):599-615.

Stapp, Henry P. (2006). "Quantum Interactive Dualism, II: The Libet and Einstein–Podolsky–Rosen Causal Anomalies." *Erkenntnis* 65 (1):117-142.

Stapp, Henry P. (2011). *Mindful Universe: Quantum Mechanics and the Participating Observer*. New York: Springer-Verlag.

Stapp, Henry P. (2014). "Mind, Brain, and Neuroscience." *Cosmos and History* 10 (1):227-231.

Stapp, Henry P. (2014a). “Quantum Physics and Philosophy of Mind.” In Uwe Meixner & Antonella Corradini (eds.), *Quantum Physics Meets the Philosophy of Mind: New Essays on the Mind-Body Relation in Quantum-Theoretical Perspective*. De Gruyter. pp. 5-16.

Stapp, Henry P. (2017). *Quantum Theory and Free Will: How Mental Intentions Translate into Bodily Actions*. Cham: Imprint: Springer.

Strawson, Galen. (1986). *Freedom and Belief*. Oxford, GB: Oxford University Press.

Strawson, Galen. (1994). “The Impossibility of Moral Responsibility.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 75: 5-24.

Strawson, Galen. (2002). “The Bounds of Freedom.” In *The Oxford Handbook of Free Will*, ed. Robert Kane, 441-460. New York: Oxford University Press.

Strawson, Peter. (1962). Freedom and Resentment. *Proceedings of the British Academy* 48:187-211.

Taylor, Richard. (1974). *Metaphysics*. Englewood Cliffs, N.J.: Prentice Hall.

Tse, Peter Ulric (2013). *The Neural Basis of Free Will: Criterial Causation*. MIT Press.

Tse, Peter Ulric. (2013a). “Free will unleashed.” *New Scientist* Issue 2920

van Inwagen, Peter. (1983). *An Essay on Free Will*. New York: Oxford University Press.

van Inwagen, Peter. (2002). “Free Will Remains a Mystery.” In *The Oxford Handbook of Free Will*, ed. Robert Kane, 158-177. New York: Oxford University Press.

van Inwagen, Peter. (2017). *Thinking About Free Will*. New York, NY, USA: Cambridge University Press.

Watson, Gary. (2003). *Free Will, 2nd Ed.* Oxford: Oxford University Press.