Chapter 10

Two Types of Libertarian Free Will are Realized in the Human Brain

Peter U. Tse

In my book *The Neural Basis of Free Will* (2013), I described various developments in neuroscience that reveal how volitional mental events can be causal within a physicalist paradigm. I began by (1) attacking the logic of Jaegwon Kim's (1993) exclusion argument (EA). According to the EA, information in general, and mental information in particular, cannot be causal, and must be epiphenomenal, because particle-level physical-on-physical causation is sufficient to account for apparent causation at all higher levels. If this is true, then mind cannot be causal in the universe. It would follow that there cannot be any free will or morality that made any difference to physical outcomes, because quark-level descriptions (or whatever is operative at the rootmost level of physical causation), where there is no need for informational, mental, or moral descriptors, would be sufficient to account for the causal unfolding of events. The first task of anyone interested in free will or mental causation, therefore must be to show where the EA breaks down. I will summarize here my past arguments that the exclusion argument falls apart if indeterminism is the case. If I am right, I must still build an account of how mental events are causal in the brain. To that end I take as my foundation (2) a new understanding of the neural code that emphasizes rapid synaptic resetting over the traditional emphasis on neural spiking. (3) Such a neural code is an instance of 'criterial causation,' which requires modifying standard interventionist conceptions of causation such as those favored by Judea Pearl (2000) and John Woodward (2003). A synaptic reweighting neural code provides (4) a physical mechanism that accomplishes downward informational causation, (5) a middle path between

determinism and randomness, and (6) a way for mind/brain events to turn out otherwise. This 'synaptic neural code' allows a constrained form of randomness parameterized by information realized in and set in synaptic weights, which in turn allows physical/informational criteria to be met in multiple possible ways when combined with an account of how randomness in the synapse is amplified to the level of randomness in spike timing. This new view of the neural code also provides (7) a way out of self-causation arguments against the possibility of mental causation. It leads to (8) an emphasis on imaginative deliberation and voluntary attentional manipulation as the core of volitional mental causation rather than, say, the correlates of the unconscious premotor computations seen in Libet's readiness potentials. And this new view of the neural code leads to (9) a new theory of the neural correlates of qualia as the 'precompiled' informational format that can be manipulated by voluntary attention, which gives qualia a causal role within a physicalist paradigm. Finally, it is not enough to simply have the 'first-order free will' afforded by the above kind of nervous system that can choose actions freely. Only if present choices can ultimately lead to a chooser becoming a new kind of chooser—that is, only if there is second-order free will or metafree will—do brains have the capacity to both have chosen otherwise, and to have meta-chosen otherwise. Only such a metafree will allows a brain to not only choose among options available now, but to cultivate and create new types of options for itself in the future that are not presently open to it. Only then can there be responsibility for having chosen to become a certain kind of person who chooses from among actions consistent with being that kind of person. In section (10) I will discuss how the brain can choose to become a new kind of brain in the future, with new choices open to it than are open to it now. In section

(11) I will argue that criterial causation gets around luck arguments against self-forming actions. I elaborate each of these ten points in the eleven sections below.[1]

1. Overturning Kim's Exclusion Argument

It is necessary to challenge this philosophical claim because if the exclusion argument is correct, a central assumption of neuroscience and psychology, namely that mental information can be causal of subsequent brain events, falls apart. A philosopher might object that Kim's argument is only an argument against anti-reductionism or non-identity theories, and that if one adopted a type identity theory, then mental information would indeed be causal, but only by virtue of being physical. But that would still make mental events, like pain, not causal in the universe by virtue of their informational characteristics, such as hurting, but only causal via their physical instantiations having physical causal efficacy. If mental events cannot be causal by virtue of being informational—and they would not be if the exclusion argument is correct—then it would follow that there can be no free will (i.e. free mental events) that makes a difference to physical outcomes. It would also follow that there can be no morality or immorality; mental decisions cannot be held accountable for their having made a difference to physical outcomes, when, as would be the case if the exclusion argument holds, mental decisions can make no difference in a universe where quark-on-quark interactions (or whatever units of energy are at the very lowest level) are sufficient to account for all causal chains. So confronting the exclusion argument is a first step to any argument for mental causation or free will, since informational or mental causation is necessary (though not sufficient for) forms of free will and moral agency that make a difference to physical outcomes.

---

[1] This chapter has been adapted from an excerpt from my upcoming book *Imagining Brains: The neural sources of human freedom and creativity*.

Let us examine Kim's (1993) exclusion argument more closely. If true, Kim's exclusion argument would logically rule out that mental information can be causal. (Note that this argument could also be used to argue that genetic information is epiphenomenal, though no one argues that, probably because we understand the genetic code quite well now, whereas we do not yet fully understand the neural code). The argument rests on a premise of the causal closure of the physical. "Causal closure" means that causality at the level of particles is *sufficient* to account for all outcomes and interactions at the level of particles.[2] Kim (2005: 17), applying Occam's razor, advocates the "exclusion of over-determination" when modeling physical causation. In his words: "If event *e* has a sufficient cause *c* at *t*, no event at *t* distinct from *c* can be a cause of *e*."

---

[2] Here is how Kim (1996: 147) defines the causal closure principle: "Pick any physical event, say, the decay of a uranium atom or the collision of two stars in distant space, and trace its causal ancestry or posterity as far as you would like; the principle of causal closure of the physical domain says that this will never take you outside the physical domain. Thus, no causal chain involving a physical event will ever cross the boundary of the physical into the nonphysical." Caruso (2012: 13) defines causal closure as follows: "If x is a physical event and y is a cause or effect of x, then y, too, must be a physical event." Note, however, that these "physical from physical" definitions of causal closure ignore the sufficiency of a cause. They hold under determinism or indeterminism as we walk causal chains backwards into the past. But as we walk causal chains into the future, which, barring backward causation in time, is the way they in fact go, under indeterminism, but not under determinism, the majority of physically possible causal chains do not happen, and therefore never become actually physical. They do not happen because, under indeterminism, a cause is not sufficient to specify which one of many possible outcomes will happen.

Note that without the sufficiency of *c*, Kim cannot apply the "exclusion of over-determination" principle, so cannot rule out mental causation. The sufficiency of *c* is crucial if the exclusion argument is to succeed at excluding mental causation. If particle-level causality is sufficient to account for particle behavior, and neurons are made of particles, mental events, assuming that they supervene on neuronal events, can play no causal role in neuronal behavior. In other words, mental events qua mental events cannot cause fundamental particles to behave differently than they otherwise would have if they had only interacted according to the laws obeyed by particles.

Put succinctly (Kim 1993: 206–210): If (i) the "realization thesis" is the case, then each mental state is synchronically determined by underlying microphysical states, and if (ii) "the causal or dynamical closure of the physical thesis" is the case, then all microphysical states are completely diachronically necessitated by antecedent microphysical states, then it follows that (iii) there is no causal work left for mental states as such to do. If the logic here is valid, then only if either (i) or (ii) is incorrect, is there potentially room to develop a theory of mental causation. So any theory of mental causation that attempts to meet "Kim's challenge" must explicitly state which premise, (i) and/or (ii), is incorrect.

If quantum domain indeterminism is correct then (ii) is incorrect, because any particular present microphysical state is not necessitated by its antecedent microphysical state or states. In other words, the traditional definition of causal closure that "every physical event has an immediately antecedent sufficient physical cause" is not satisfied, because when a cause *c* can be indeterministically followed by any number of possible effects e_i, then *c* is not a sufficient cause of any of the possible e_i, because they might not happen if they have not yet happened, and they might not have happened even after they have happened. Papineau (2009) tries to handle the problem of causal non-sufficiency of *c* introduced by indeterminism by appending a qualifier to

the more traditional definition of causal closure as follows: “Every physical effect has an immediate sufficient physical cause, in so far as it has a sufficient physical cause at all.” A similar attempt to make—in this case Davidson’s—definition of causal closure consistent with indeterminism is to say that ‘every physical event *that has an explanation* has a physical explanation.’ But neither of these attempts to dodge the non-sufficiency of *c* imposed by indeterminism gives existing physical explanations enough credit. Quantum-domain effects are not unexplained. It is not the case that just anything can happen inexplicably. Rather, the set of possible outcomes and their likelihoods of occurrence are very precisely defined by quantum theory, arguably the most accurately predictive theory in the history of science.

Classical deterministic laws are laws that hold among sufficiently causal actualia, where both *c* at *t1* and *e* at *t2* are actual events. Quantum mechanical laws are deterministic at the level of possibilia, but indeterministic at the level of actualia, because which possible outcome will occur upon measurement is only probabilistically specifiable. Nonetheless, under quantum mechanics *c* is sufficiently causal of its entire set of possible outcomes e_i with their associated probabilities of occurring. It is just that *c* is not a sufficient cause of any particular one of its many possible effects that happens to happen when measured. Classical deterministic and modern quantum mechanical laws both operate deterministically, and causation is sufficient, but over different types of physical entities: actualia and possibilia, respectively. Actualia and possibilia, while both physical, have mutually exclusive properties. Actualia are real and exist now or in some past moment; they have a probability = 1 of happening or having happened. Possibilia are not yet real and may never become real, and exist in the future relative to some *c*, and have a probability of happening between zero and one. A given event cannot be both actual and possible at the same time.

Closure, therefore, applies to different types of physical events under ontological determinism and indeterminism. "Closure" entails that the set of physical events is closed; any particular effect will be a member of the same set to which a sufficient cause itself belongs. Determinism is closed at the level of actualia; any particular cause or effect will be a member of the set of all actual events in the universe across all time. Indeterminism, in contrast, is not closed at the level of actualia. This is because a non-sufficient actual cause and one of its possible outcomes that may never happen are not both members of the set of actualia. Rather, quantum theory is closed (and deterministic!) at the level of possibilia: Any particular outcome or event will be a member of the set of all possible outcomes or events in the universe across all time, and any possible cause is sufficient to account for the set of all of its possible effects. Under indeterminism physical explanations are of a different type than under determinism, though both actualia and possibilia are physical, and theories of either actualia or possibilia are physical explanations.

An indeterministic causal closure thesis could be restated as follows: "(ii*) the set of all possible microphysical states is completely diachronically necessitated by antecedent possible microphysical states." The realization thesis for the indeterministic case might be: "(i*) all mental states are synchronically determined by underlying sets of possible microphysical states." But claim (i*) is contrary to the definition of supervenience. Mental events do not supervene on sets of possible physical states, they supervene on specific, actually occurring physical states. Since it is absurd to maintain that mental events synchronically supervene on sets of possibilia, we can rule (i*) out. It remains to be shown whether (i), i.e. supervenience on actualia, can be combined with (ii*), i.e. causal sufficiency and closure among possibilia, to yield (iii). We will see below that this combination fails to deliver causal closure.

An actual microphysical state and the set of all possible microphysical states are different kinds with mutually exclusive properties (e.g., real/~real; present/~present). The essentially syllogistic structure of the exclusion argument requires staying within a logical kind. It is logically valid to draw from the major premise (ii) 'All physical events are caused by preceding sufficient physical causes' and the minor premise (i) 'mental events are realized in physical events' the conclusion (iii) that 'the physical events that realize mental events have preceding sufficient physical causes'. But now we are splitting 'physical' into two types with mutually exclusive properties, possibilia and actualia. The conclusion (iii) of the syllogism holds only if both the major and minor premises hold and are both about actualia as in (ii) and (i), or both are about possibilia as in (ii*) and (i*). If one premise is about possibilia and the other about actualia, the conclusion does not follow, because the premises are about exclusive entities. For example, (ii) and (i*) would read 'All actual physical events are caused by preceding sufficient actual physical causes' and 'mental events are realized in sets of possible physical events,' which violates syllogistic logic as much as 'all men are mortal' and 'Socrates is a robot.' Conversely, (ii*) and (i) would read 'The set of possible physical events are caused by preceding sufficient possible physical causes' and 'mental events are realized in actual physical events,' which similarly violates syllogistic logical form. Thus, assuming indeterminism, mental causation is not logically ruled out by Kim's argument.[3]

---

[3] As an aside, there is another argument that (i) with (ii*) cannot logically entail (iii). Obviously, causes must precede effects. The usual exclusion argument is that (ii) diachronic actual elementary particle interactions *preceding* the moment *t* of (i) synchronic mental supervenience on actual particle configuration *p* leaves no room for mental events qua mental events to have any causal effect since those preceding physical interactions are sufficient to cause *p*. However,

I wrote a version of the above argument first in my 2013 book, and then on the philosophy blog *Flickers of Freedom*, where I battled with free will denier Neil Levy. He wrote

---

if (ii*) is taken to refer to a diachronic set of possible events preceding (i) mental supervenience on *p*, then there is a problem, because possibilia do not exist in the past of *p*; only actual events, such as those described in (ii) do. Once we have reached time *t* and *p* is not a possibility but an actuality, then all events prior to *t* must also be actual; events in the past are actual events that happened and are no longer possible. If they were possible they would lie in the future. Possibilia only exist in the future relative to some actual or possible event. But *p* we agree is actual since supervenience makes no sense for possibilia, as in (i*), which we have rejected. Alternatively, if we want to think of the possibilia in (ii*) 'collapsing' into *p*, where *p* was one among many possibilities, much like the quantum mechanical collapse of the wave function, we are again left with the problem that the set of possibilia is not sufficient to cause *p* per se, because *p* might not have happened at all and some other possible outcome might instead have happened. However, if the possibilia in (ii*) are taken to temporally follow (i) the actual *p* at *t*, well, that is certainly consistent with the idea that possibilia can exist in the future of *p*. But then possibilia in the future of *p* would be seen as being sufficiently causal of *p*, which would entail impossible backward causation in time. Thus the possibilia described in (ii*) can neither precede nor follow the actualia described in (i) and be sufficiently causal of them. In sum, (i) and (ii*) do not together entail (iii), whether on logical (syllogistic) grounds or on the grounds that possibilities can only exist in the future and not in the past of actual events such as those on which mental events supervene. Again, assuming indeterminism, mental causation is not logically ruled out by Kim's argument.

that the above argument "…is badly confused. It rests on a misunderstanding regarding the causal closure principle. Tse understands the principle to claim that physical causes are sufficient for the occurrence of physical effects. If indeterminism is true, then physical causes sometimes or often are not sufficient for the occurrence of later events. Tse therefore concludes that the closure principle is false for indeterministic systems, so it is no obstacle to mental causation. But the causal closure principle is, roughly, the principle that physical events can be accounted for by physical causes, or (equivalently) that physics is causally complete. It is silent on whether physics is deterministic or not. The brain may be indeterministic; causal closure remains an obstacle to mental causation."[4]

In response to Levy, I did not invent the definition of causal closure as "every physical effect having an immediately antecedent *sufficient* physical cause"; many philosophers have written variants of just such a definition, including Papineau and Kim, cited above. If we eliminate the requirement that *c* be sufficient to cause its physical effects, we lose Kim's elegant "exclusion of over-determination" argument against any possible causal role of the mental qua mental, and can no longer rule that out. In the absence of sufficient physical causation we could at most argue that an action or outcome would be overdetermined if it has both a physical cause (whether deterministic or indeterministic) and a mental cause. Under that move all the causal

---

[4] I find the views of Levy and others who deny mental causation and free will to be wrong, nihilistic and impoverished; wrong for reasons covered in my book and here, nihilistic because there can be no moral responsibility or self forming acts under such views, and impoverished because they fail to recognize the astounding elaboration of modes of top-down informational causation that have evolved in biological systems, including principally the causal roles of our minds in realizing our own envisioned futures.

work is still done by, presumably, particle on particle interactions, not by mental events qua mental events (e.g. when pain, by virtue of consciously hurting, causes a trip to the dentist). As a physicalist I agree that "physical events can be accounted for by physical causes." But there is an ambiguity in Levy's phrase "accounted for" here. Deterministic physical laws account non-probabalistically (or rather, with a probability of 1) for a deterministic succession among actualia, whereas indeterministic physical laws account probabilistically for an indeterministic succession among actualia; or, as is the case with the evolution of the wave function in quantum mechanics, physical laws account deterministically for a changing probability distribution of possible outcomes of measurements. If we are to take the idea of closure of the physical seriously, then a physical cause *c* and its physical effect(s) must belong to the same closed set. We agree that this closed set includes only physical events whether determinism or indeterminism is the case. But under determinism that closed set of physical events includes physical actualia across time whereas under indeterminism it includes physical possibilia across time. In principle, classical physics is a causally complete and deterministic account of the sequence of actualia over time, and quantum physics is a causally complete and deterministic account of the sequence of possibilia over time. But standard versions of quantum physics do not give a complete account that can explain why one possible outcome becomes actual upon measurement or observation rather than other possible outcomes that did not occur. It just happens, with no reason given beyond chance. If *c* does not provide sufficient grounds for why one possible outcome occurs over another, exclusion of overdetermination cannot be used to rule out the possibility that the physical realization of present mental events might bias which particle possibilia will become actualia in the imminent future. Note that this does not require positing any bizarre notions like consciousness collapsing the wave packet. It just requires that present

physically realized informational criteria placed on inputs can be met in the future in multiple possible ways.

In sum, Kim's exclusion argument amounts to saying that the physical substrate does all the causal work that the supervenient mental state is supposed to do, so mental or informational events can play no causal role in material events. One might say that this does not hold if the mental and physical are identical, but even then it is the physical side of the equation where causal efficacy resides. On Kim's reductionistic view, all causation "seeps away," as Ned Block put it, to the rootmost physical level, i.e. particles or strings or whatever physicists next model the most basic level to be like. Add to that an assumption of determinism, and the laws of physics applicable at the rootmost level are sufficient to account for event outcomes at that level and every level that might supervene on that level. So informational causation, including voluntary mental causation or any type of libertarian free will that relies on information being causal in this universe, is ruled out. I argue that indeterminism undermines this sufficiency, so provides an opening whereby physically realized mental events could be downwardly causal.

Exploiting this opening, biological physical systems evolved to emphasize a new kind of physical causation, one based upon triggering physical actions when detected spatiotemporal patterns in energy meet the criteria for triggering. This is a very different kind of causation than traditional Newtonian conceptions of the causal attributes of energy, such as mass, momentum, frequency or position, which seem to underlie deterministic and exclusionary intuitions. But patterns, unlike amounts of energy, lack mass and momentum and can be created and destroyed. They only become causal if there are physically realized pattern detectors that respond to some pattern in their energetic inputs. Basing causal chains upon successions of detected patterns in energy, rather than the transfer of energy among particles, opens the door not only to

informational downward causation but to causal chains (such as mental causal chains or causal chains that might underlie a game of baseball or poker) that are not describable by or solely explainable by the laws of physics applicable at the rootmost level. Yes, a succession of patterns must be realized in a physical causal chain that is consistent with the laws of physics, but many other possible causal chains that are also consistent with physical laws are ruled out by informational criteria imposed on indeterministic particle outcomes. Physically realized informational criteria set in synaptic weights effectively sculpt informational causal chains out of the 'substrate' of possible physical causal chains. Information is not causal as a force. Rather it is causal by allowing only those possible physical causal chains that are *also* informational causal chains (i.e. that meet particular preset informational criteria) to become real (i.e., to switch ontological status from possibilia to actualia).

The old argument that there is no middle ground between utter randomness and determinism is wrong. If indeterminism is ontologically the case, then parameters placed on possible outcomes can select from among possible particle paths just that subset that also satisfies specified informational parameters. Causation via informational reparameterization would not be possible if the neural code were based on spikes ballistically triggering spikes like Newtonian billiard balls deterministically triggering the motions of the billiard balls that they collide with. But if presynaptic neural spikes reparameterize the informational criteria that will make postsynaptic neurons spike given possible future presynaptic neural spike inputs, then many neural causal chains are possible that would be consistent with those reset informational parameters or criteria for firing. By itself neural causation via informational reparametrization does not get us the control or the ability to settle outcomes that is needed for free will and moral responsibility. Much more is needed and has indeed evolved to be present in our brains, which I

will return to later. But an informational criterial neural code is a necessary condition for having such control. For example, if I say to you "Name a female politician with red hair" your response will likely not be utterly random, because you will state a name that meets these three criteria of being a woman, a politician and having red hair. But your response is also not determined, because your answer might have turned out otherwise. For example, if you responded "Angela Merkel," had I been able to rerun the universe again from the moment of my question, this time you might have said "Margaret Thatcher." This is because the brain has in fact evolved to amplify quantum domain randomness, as I go into in depth in my book, up to a level of neural spike timing randomness. And since neurons are effectively spike coincidence detectors, this randomness affords the possibility of other solutions to any given finite set of informational critieria. This kind of criterial neural code in turn affords the possibility that events might turn out otherwise, yet not be utterly random, because they will have to meet the informational criteria that were preset. Information, then, is not causal as a force again, but more as a filter that allows possibilia (at the particle level) that are consistent with informational parameters to become actualia, and those that are not consistent with informational parameters to get weeded away. The brain will need more causal powers to get to a full-blown libertarian free will, which have also evolved, as I will argue later. But a criterial or parametric neural code is necessary (even when not sufficient) for free will and moral responsibility, because informational reparameterization via synaptic weight resetting is the core engine whereby information can be causal of subsequent events in the brain. Thus all possible or actual informational causal chains are also possible or actual physical causal chains, whereas the vast majority of possible physical causal chains are not informational causal chains. Only those who have yet to appreciate that causation in the brain can proceed via informational criterial or parameter resetting via rapid

synaptic weight changes can continue to bring out the tired Humean argument that there can be no libertarian free will realized in the brain because there is nothing between determinism (where events could not have turned out otherwise) and utter randomness (where an agent plays no role in the chance events that happen next).

Accounts of free will that require supernatural or contra-causal interventions have given libertarianism a bad name, and violate basic assumptions of physicalism and science. For any naturalistic variant of libertarian free will to exist, several necessary conditions must be met. First, indeterminism must be ontologically real, rather than just a matter of epistemic uncertainty. Under indeterminism, I argued above, Kim's exclusion argument fails to rule out mental causation, leaving us with an opening to develop a believable account of mental or informational causation that is not epiphenomenal. To get there, the following facts must in turn be true of neural processing: (1) quantum domain randomness must be amplified up to the level of randomness in macroscopic neural information processing, which (2) would have to be able to harness this randomness to fulfill information processing aims, and (3) there would have to be a role for the subclass of information processing that we call 'mental,' particularly the subclass of the mental that we regard as consciously volitional, in the specification of the ends to which such harnessing will apply, if conscious willing is to be agentic or causal of the realization of such aims. In my book I laid out a detailed case that these conditions are met, permitting the physical realization of a 'type-1 libertarian free will.' In order to have a 'type-2 libertarian free will,' however, an additional condition would have to be met, namely, the nervous system would have to (5) be able to make decisions about how it would like to change itself, and then have the means to change itself, over time, into the intended type of decider.

Libertarian free will requires non-illusory downward mental causation. ‘Downward’ here means that events at a supervening level can influence outcomes at the rootmost level. In this context it would mean that information can bias which possible particle paths are realized. There is no wiggling out of this. If we want mental causation, and a free will and moral responsibility rooted in mental events that cause real consequences, we must defend the position that an informational entity, such as an intention or plan developed or held in working memory, can bias what possible particle paths open at the rootmost level can and do become real. But an entity at a supervening level cannot, logically, change its own physical basis because there can be no *causa sui*. This is where criterial causation via informational reparameterization comes in, because this allows what supervenes now to place informational constraints on what can supervene in the future.

How might such constraining work in the brain? The key pattern to which neurons respond is temporal coincidence. A neuron will only fire if it receives a certain number of coincident inputs from other neurons. Criterial causation occurs where physical criteria imposed by synaptic weights on coincident inputs in turn realize informational criteria for firing. This permits information to be downwardly causal regarding which indeterministic events at the rootmost level will be realized; only those possible rootmost physical causal chains that meet physically realized informational criteria can drive a postsynaptic neuron to fire, and thus become causal at the level of information processing. Typically the only thing that the set of all possible rootmost physical causal chains that meet those criteria have in common is that they meet the informational criteria set. To try to cut information out of the causal picture here is a mistake; the only way to understand why it is that just this subset of possible physical causal

chains—namely those that are also informational causal chains—can occur, is to understand that *informational* criteria delimit that class of possible outcomes.

As Eddy Nachmias put it on my Oct. 2013 thread on the blog *Flickers of Freedom*: "the fact that informational state S1 could be realized by a range of physical states P1-PN and that informational state S2 counterfactually depends on S1 but *not* any one of the specific physical states, including the one that actually realizes S1 on this occasion (e.g., P3) suggests that S1 is what makes a difference to S2 in a way that P3 does not. If we want to causally manipulate S2, manipulating P3 may not do it (e.g., if we alter it to P1 or P4, or one of the other S1 realizers); rather, we need to manipulate S1 (yes, by altering its realizers in the right way, but the right way will involve considerations of the S-level, not the P-level). S2 *rather than* S7 occurs *because* S1 *rather than* S1' occurred, and not because P1 rather than P4 occurred."

Information only comes into existence by virtue of a decoder receiving input that matches its conditions (typically placed on the phase relationships or patterns in incoming energy) for the release of some effect, say, an action potential sent to other such decoders. But a decoder also serves as a 'filter' on the set of all potentially causal inputs, since it will only change the system of decoders in which it is embedded, namely, by firing, if its physically realized informational criteria are met.

Information cannot be anything like an energy that imposes forces, because it is not material even when it is realized in the material substrate. Information's causal power consists in 'filtering' informational causal chains out of the set of all possible physical causal chains by constraining which sets of possible physical causal chains can occur. Although every informational causal chain is also a physical causal chain, most physical causal chains are not informational causal chains. Information is downwardly causal not as a material force, but as

constraints that only allow the realization of sets of possible physical causal chains at the rootmost level that also comprise informational causal chains. Physical laws are not violated by this. Every possible physical causal chain conserves energy and momentum and so forth. But only those possibilities allowed by physical laws which *also* meet informational criteria pass the physically realized informational filter, and become informationally causal, either by reparameterizing the criteria by which other neurons will assess future input, namely, by changing their synaptic weights, or by triggering other neural firing.

Information is multiply realizable because which particular set of spike inputs—and thus what particular information—will make the neuron fire is unforeseeable, so long as the physical/informational criteria for firing are met. If neural causal chains are also informational causal chains, and informationally equivalent informational causal chains are realizable in multiple different neural or particle causal chains, then the parsimonious model is one of information causing information. Yes, there must always be some physical realization of information, but under physical/informational criterial causation, which one it happens to be is irrelevant so long as informational criteria are met. Chains of successive informational criterial satisfactions and criterial resettings afford the physical realization of downward mental causation.

On the same *Flickers of Freedom* blog Derk Pereboom said, "if on some proposal, a dualist or a nonreductivist one, M and P are distinct causes of E, the threat posed by exclusionary reasoning will be neutralized by any response on which the number of causes is reduced to just one. There are two ways to achieve this: a first is by eliminating all but one of the causes, and the second is by identifying the causes." If mental events are a type of information, and information is identical to acts of decoding immaterial (i.e. not made of mass) relationships or patterns

among physical inputs, then mental events are identical to some class of acts of decoding. But note, this identity does not make mental events have physical properties like mass or momentum, because the identity is not with physical events at some instant, but with a process that is realized in physical events. Moreover, acts of decoding patterns cannot be reduced to a level where the patterns are not explicit, say the rootmost level, because the decoder only responds to a pattern at a level where it is explicit. And at that level, potentially countless configurations of rootmost events are equivalent in that they each realize the same pattern as far as the decoder is concerned. Thus the identification is not with events at the microscopic level or even the neuronal level, but at the level of decoding the non-physical patterns to which the decoder is sensitive. Under determinism supervening informational criteria cannot filter out possible but non-informational causal chains at the rootmost level, because there is only one possible causal chain. But if indeterminism is the case, supervening informational criteria can make a difference regarding which possibilities at the rootmost level happen. That is, under indeterminism but not determinism, there is non-redundant causal work for informational criteria to do.

But how does this give the brain the capacity to freely will? It is not enough for neurons to filter out non-informational possible physical causal chains. It must be the case that some neural activity that we associate with volition can control the parameters that neurons will apply in the future to enact such acts of filtering. Control comes from executive circuits that can plan, imagine, deliberate and make decisions in light of highest level demands and needs, and that can 'rewire' circuits and reparameterize neurons, by changing synaptic weights, to embody new informational criteria for firing that will fulfill current executive ends. The downward causation afforded by the informational filtering of possible rootmost causal chains becomes agentic

downward causation when executive circuits can rewire lower level circuits to fulfill whatever criteria they demand.

2. A New View of the Neural Code

How might informational reparameterization and causation work at a neural level? In my 2013 book I developed a new understanding of the neural code that emphasizes rapid and dynamic synaptic weight resetting over neural firing as the core engine of information processing in the brain. The neural code on this view is not solely a spike code, but a code whereby information is transmitted and transformed by flexibly and temporarily changing synaptic weights on a millisecond timescale. One metaphor is the rapid reshaping of the mouth (analogous to rapid, temporary synaptic weight resetting) that must take place just before vibrating air (analogous to spike trains) passes through, if information is to be realized and communicated. What rapid synaptic resetting allows is a moment by moment changing of the physical and informational parameters or criteria that have to be met before a neuron will fire. This dictates what information neurons will be responsive to and what they will 'say' to one another from moment to moment. Thus the heart of criterial causation in the brain is the resetting, by other neural inputs, of the synaptic weights that realize informational parameters that have to be met by a neuron's subsequent inputs in order for that neuron to fire, which in turn will reset the parameters that will make other neurons subsequently fire.

3. Rethinking Interventionist Models of Causation

Interventionist/Manipulationist models of causation (e.g. Pearl 2000; Woodward 2003) are rooted in the intuition that if some event A causes some event B, then one should be able to manipulate A in some way and see corresponding changes in B after changing A. If A is modeled as causing B, then there should be an intervention on A (in at least some state of the

model) that results in B changing its value. These kinds of models of causation basically describe what scientists already do to determine causal relationships among variables. Scientists have for centuries tried to control for all independent variables (Woodward calls this "screening off" the other variables besides A that likely partially cause B, by holding their values constant) save one, A, which they vary, in order to see the consequences or changes expressed by some outcome or dependent variable B. If B changes with an intervention on A, it is concluded that A, among perhaps other causal variables, in part causes B.

A counterfactual formulation of interventionism would be: "If A had not occurred, with all screened off variables that may cause B held constant, then B would not have occurred." A core point of criterial causation is that we need to enhance the interventionist account by saying: "If A had not occurred, with all screened off variables that may cause B held constant, *and with the parameters by which B evaluates its inputs also held constant*, then B would not have occurred."

Standard interventionist models of causation carry out some intervention on A to determine what effects, if any, there might be on B (and other variables). If instead of manipulating A, or A's output to B, however, we instead manipulate the criteria, parameters or conditions that B places on its input (including on input from A), which must be satisfied before B changes or acts, then changes in B do not follow passively from changes in A as they would if A and B were, say, billiard balls. Inputs from A can be identical, but in one case B changes in response to A, and in another case it does not, depending on B's criteria for responding. This reparameterization of B is what neurons do when they change each other's synaptic weights, such that a neuron now responds optimally to different inputs than prior to the act of physical and informational reparameterization or criterial resetting. Criterial causation emphasizes that

what can vary is either outputs from A to other nodes (the traditional, and I would say incomplete view of causation), or how inputs are decoded by receiving nodes B, B', B'' and so on. On this view, standard interventionist (hearkening all the way back to 'Newtonian' models of causation that emphasize energy transfer and conservation; e.g. P. Dowe's views (1992)) are a special case where B places no particular conditions on input from A that have to be met before B changes state. But the brain, if anything, emphasizes causation via reparameterization of B, by, for example, rapidly changing synaptic weights on post-synaptic neurons. Let me emphasize that I do not think that Woodward or Pearl are wrong. But they also make no mention, as far as I can tell, that causation might partly depend on reparameterizations of B. Thus their views of causation were incomplete and need to be amended by emphasizing the role of informational reparameterization of the response characteristics of B.

Changing the code or parameters or criteria that B uses to decode, interpret or respond to input is a manipulation that might make no apparent changes to A or any other variable in the system for long and uncertain spans of time, until just the right pattern of inputs comes along. This is quite different from the unamended traditional manipulationist view, where a manipulation of A is expected to alter B within a short duration that is dictated by whatever physical laws are thought to apply, and certainly not after unspecifiably long durations, as is the case under my amended view. For example, the Mossad might program a cellphone to explode only when a particular phone number, known 'only' to their target, is dialed. Manipulating 'A' here appears to have no effect on any dependent variable 'B' and might not, in principle, for as long as you like (think of a booby trap in a king's tomb set by the pharaoh's builders that only kills archaeologists millennia later). It might take years to work this phone up the ranks and into the hands of their targeted Hamas leader. But when the bomb maker dials his 'secret' number to

call his uncle in Paris, his head is blown off. This kind of reparameterization of B need not have immediate noticeable or measurable effects within the system, so seems to violate the assumption of the traditional view that causation is transferred at some fast speed (say the speed of light). But reparameterization of B is a causal intervention nonetheless, even though this subclass of causation has been relatively ignored by philosophers so far. It is at the heart of what I mean by 'criterial causation.' Other names for this might be 'reparameterization causation,' 'pattern causation' or 'phase causation.'

I believe, however, that it was the 'discovery' of this class of causation by evolution that led to the explosion of physical systems that we now call biological systems. Once causation by reparameterization came not only to involve conditions placed on physical parameters (e.g. molecular shape of a neurotransmitter before an ion channel would open in a cell membrane), but also conditions placed on informational parameters (fire above or below baseline firing rate if and only if the criteria for a face are met in the input) that were realized in physical parameters (fire if and only if the criteria on the simultaneity of spike inputs are met), we witnessed a further revolution in natural causation. This was the revolutionary emergence of mind and informational causation in the universe, as far as we know, uniquely on Earth, and perhaps for the first time in the history of the universe.

4. How Downward Causation Works

Downward causation means that events at a supervening level can influence outcomes at the rootmost level. In this context it would mean that information could influence which possible particle paths are actualized. While it would be impossible self-causation if a supervening event changed its own present physical basis, it is not impossible that supervening events, such as mental information, could bias future particle paths. How might this work in the brain? The key

pattern in the brain to which neurons respond is temporal coincidence of arriving action potentials from other neurons. A neuron will only fire if it receives a certain number of coincident inputs from other neurons. Criterial causation occurs where physical criteria imposed by synaptic weights on coincident inputs in turn realize informational criteria for firing. This permits information to be downwardly causal regarding which indeterministic events at the rootmost level will be realized; only those rootmost physical causal chains that meet physically realized informational criteria can drive a postsynaptic neuron to fire, and thus become causal at the level of information processing. Typically the only thing that the set of all possible rootmost physical causal chains that meet those criteria have in common is that they meet the informational criteria set. To try to cut information out of the causal picture here is a common but serious mistake; the only way to understand why it is that just this subset of possible physical causal chains—namely those that are also informational causal chains—can occur, is to understand that it is informational criteria that dictate that class of possible outcomes.

The information that will be realized when a neuron's criteria for firing have been met is already implicit in the set of synaptic weights that impose physical criteria for firing that in turn realize informational criteria for firing. That is, the information is already implicit in these weights before any inputs arrive, just as what sound your mouth will make is implicit in its shape before vibrating air is passed through it. Assuming indeterminism, many combinations of possible particle paths can satisfy given physical criteria, and many more cannot. The subset that can satisfy the physical criteria needed to make a neuron fire is also the subset that can satisfy the informational criteria for firing (such as 'is a face') that those synaptic weights realize. So sets of possible paths that are open to indeterministic elementary particles which do not also realize an informational causal chain are in essence "deselected" by synaptic settings by virtue of the

failure of those sets of paths to meet physical/informational criteria for the release of a neural spike. A neural code based on informational reparameterization of subsequent neural firing affords the possiblity of top-down causation because an informational command such as “think of a woman politician with red hair,” whether externally heard or internally generated by executive processes, can reparameterize subsequent physical neural activity such that the result is, randomly within those parameters, Angela Merkel or, equivalently, Margaret Thatcher.

5. Between Determinism and Randomness

Let us return to Hume. Way back in 1739 he wrote “‘tis impossible to admit of any medium betwixt chance and an absolute necessity.” Many other philosophers have seen no middle path to free will between the equally ‘unfree’ extremes of determinism and randomness. They have either concluded that free will does not exist, or tried to argue that a weak version of free will, namely, ‘freedom from coercion,’ is compatible with determinism. A weak free will, where events could not have turned out otherwise than they were destined to turn out since the beginning of a deterministic universe, is by definition compatible with determinism in that our determined decisions are uncoerced while certainly playing a causal role in our subsequent actions. But a strong free will, where events could have turned out otherwise, is incompatible with determinism, because only given an ontological indeterminism can events really have turned out otherwise than they did. Indeterminism is a necessary condition of a strong free will, such as a type 1 libertarian free will, or a stronger free will, such as a type 2 libertarian free will or metafree will. Thus compatibilism holds regarding weak free will while incompatibilism holds regarding strong free will and metafree will. A lot of confusion occurs because compatibilists and incompatibilists are talking past each other, assuming conflicting notions of free will.

A Humean freedom from coercion offers a ‘weak’ conception of free will that is by definition compatible with determinism, since no mention is made of a need for outcomes to have the possibility of having turned out otherwise. A libertarian conception of free will, according to which events really might have turned out otherwise, however, is not compatible with either determined or random choices, because in the determined case there are no alternative outcomes so events could never turn out otherwise, while in the random case what happens does not happen because it was willed. A libertarian free will requires meeting four high demands: Beings with strong free will (1) must have information processing circuits that have multiple courses of physical or mental activity open to them; (2) they must really be able to choose among them; (3) they must be or must have been able to have chosen otherwise once they have chosen; and (4) the choice must not be dictated by randomness alone, but by the informational parameters realized in those circuits. (Although an agent is not needed at the stage that settles which option will happen, one is needed at the stage that settles that this set of criteria will be set rather than others). This is a tough bill to fill, since it seems to require that acts of free will involve acts of self-causation. I argue that these conditions for a libertarian free will are realized in the nervous system. We have no choice but to have a libertarian free will, because evolution fashioned our nervous systems to have it. Those animals that had a nervous system that realized a libertarian free will survived to the point of procreation better than those that did not.

Criterial causation offers a middle path between the two extremes of determinism and randomness that Hume was not in a position to see, namely, that physically realized informational criteria parameterize what class of neural activity can be causal of subsequent neural events. The information that meets preset physical/informational criteria may be random to a degree, but it must meet those preset informational criteria if it is to lead to neural firing, so

is not utterly random. Preceding brain activity specifies the range of possible random outcomes to include only those that meet preset informational criteria for firing. Thus volitionally present informational parameterization of future firing is causal in the universe because it is a special subclass of that information that is causal in the universe. Such information is causal in the universe, and not epiphenomenal, by virtue of allowing only that subset of possible futures open at the particle level to become real which also realize informational causal chains. These are those that are consistent with the informational parameters or criteria that were preset by prior neural firing in present neural synaptic weights that realize informational constraints on allowable triggers of future neural firing.

6. How Brain/Mind Events Can Turn Out Otherwise

The key mechanism, I argue, whereby atomic level indeterminism has its effects on macroscopic neural behavior is that it introduces randomness in spike timing. There is no need for bizarre notions such as consciousness collapsing wave packets or any other strange quantum effects beyond this. For example, as described in detail in my 2013 book, quantum level noise expressed at the level of individual atoms, such as the single magnesium atoms that block NMDA receptors, is amplified to the level of randomness and near chaos (criticality domain) in neural and neural circuit spiking behavior. A single photon can even trigger neural firing in the retina in a stunning example of amplification from the quantum to macroscopic domains. The brain evolved to harness such ‘noise’ for information processing ends. Since the system is organized around coincidence detection, where spike coincidences (simultaneous arrival of spikes) are key triggers of informational realization (i.e. making neurons fire that are tuned to particular informational criteria), randomizing which incoming spike coincidences might meet a neuron’s criteria for firing means informational parameters can be met in multiple ways just by chance.

7. Skirting Self-Causation

A synaptic account of the neural code also gets around some thorny problems of self-causation that have been used to argue against the possibility of mental causation. The traditional argument is that a mental event realized in neural event x cannot change x because this would entail impossible self-causation. Criterial causation gets around this 'no *causa sui* argument' by granting that present self-causation is impossible. But it allows neurons to alter the physical realization of possible *future* mental events in a way that escapes the problem of self-causation of the mental upon the physical. Mental causation is crucially about setting synaptic weights. These serve as the physical grounds for the informational parameters that must be met by unpredictable future mental events realized in unpredictable future spike inputs to a neuron that will fire or not, depending on whether those physically realized informational parameters were met or not.

8. Voluntary Attention and Free Will

I argue that the core circuits underlying free choice involve frontoparietal and default mode circuits that facilitate deliberation among options that are represented and manipulated in executive working memory areas. Playing out scenarios internally as virtual experience allows a superthreshold option to be chosen before specific motoric actions are planned. The chosen option can best meet criteria held in working memory, constrained by conditions of various evaluative circuits, including reward, emotional and cognitive circuits. This process also harnesses synaptic and ultimately atomic level randomness to foster the generation of novel and unforeseeable satisfactions of those criteria. Once criteria are met, executive circuits can alter synaptic weights on other circuits that will implement a planned operation or action. For example, someone, say one of the Wright brothers, can imagine different flying machines, and then, after much deliberation, go and build an airplane that will transform the physical universe.

9. A New Theory of Qualia

Paradigmatic cases of volitional mental control of behavior include voluntary attentional manipulation of representations in working memory and the voluntary attentional tracking of one or a few objects among numerous otherwise identical perceived objects. If there is a flock of indistinguishable birds, for example, there is nothing about any individual bird that makes it more salient. But with volitional attention, any bird can be marked and kept track of. This salience is not driven by anything in the stimulus. It is voluntarily imposed on bottom-up information, and can lead to eventual motoric acts, such as shooting or pointing at the tracked bird. This leads to viewing the neural basis of attention and consciousness as not only realized in part in rapid synaptic reweighting, but also in particular patterns of spikes that serve as higher level units that traverse neural circuits (aside for neuroscientists: what I call the 'NMDA channel of communication,' commonly associated with gamma and high-gamma power in electroencephelogram data). Qualia are necessary for volitional mental causation because they are the only informational format available to volitional attentional operations. Actions that follow volitional attentional operations, such as volitional tracking, cannot happen without consciousness. Qualia on this account are a 'precompiled' informational format made available to attentional selection and operations by earlier, unconscious information processing. The relationships between qualia, free will, working memory and volitional attention are therefore very intimate. The domain of qualia is the domain of representations that either are now being volitionally attended or that could be so attended in the next moment in light of current criteria held in working memory.

10. The Human Brain Realizes Metafree Will

Even a tiger would have the kind of 'first-order or type-1 libertarian free will' afforded by the kind of nervous system summarized in the above points, in that a tiger can choose among considered actions freely, and those choices/actions could have turned out otherwise. But only if present choices can ultimately lead to a chooser turning into a new kind of chooser—that is, only if there is second-order or type-2 libertarian free will or a metafree will—do brains have the capacity to both have chosen otherwise, and to have meta-chosen otherwise. Only such a metafree will allows a brain to not only choose among options available to it now, but to cultivate and create new types of options for itself in the future that are not presently open to it. Only then can there be responsibility for having chosen to become a certain kind of person who chooses from among actions consistent with being that kind of person. Thus, in addition to meeting the four conditions that must be met by a strong free will or first-order libertarian free will, discussed above, a second-order or type-2 libertarian free will, or metafree will, must meet an additional fifth condition: (5) present choices must trigger actions that, after perhaps long durations of training or practice, ultimately lead to the reconformation of the nervous system such that a brain or chooser can decide to train itself to become a new kind of brain or chooser in the future. The human brain can choose to become a new kind of brain in the future, with new choices open to it then that may not be open to it now. This is possible because of a slower kind of neural plasticity, rooted in long-term potentiation of synaptic weights that lead to the reconformation of neural circuits. For example, one can choose to learn Chinese, and then, within a year become a brain that can adequately process and produce Chinese inputs. A tiger might have type-1 libertarian free will, but lacks type-2. No tiger thinks to itself, "next year I would like to be a different kind of tiger." This is why animals are amoral, despite having type-1 freedom of choice and action. Unlike humans, though they can choose, they cannot choose to

become a new kind of chooser. Humans, in contrast, bear a degree of responsibility for having chosen to become the kind of chooser who they now are.

Incompatibilists like Kane write about two different kinds of freedom of will, both of which are incompatible with determinism by definition. The first type of libertarian free will requires that one could have done otherwise. This is by definition not possible under determinism barring exotic philosophical moves like granting agents the capacity to change the past or the laws of physics, which most scientists, I suspect, would argue no one could actually accomplish. Choosers can only have chosen otherwise if some physical events themselves could have turned out otherwise. This makes indeterminism a necessary condition for the truth of incompatibilism. But the reality of indeterminism is not sufficient, because indeterminism might turn out to be true, yet an agent could lack the freedom to shape future events; unless the agent plays some role in defining which chance events can happen, chance just happens to the agent randomly. The freedom afforded by a nervous system that met the four conditions of a libertarian free will, such as might be realized in the brain of a tiger or other non-human vertebrate, would be what Kane might call 'freedom of action.' But this would not be what Kane would mean by a truly free libertarian free will. For that an agent would need type-2 incompatibilist free will or metafree will: the capacity to choose to become a new kind of chooser in the future. According to this notion of metafree will, the agent must have freedom of action that allows the agentic shaping of the nature of future volitional decisions. In effect, the agent must be able to shape the basis or grounds of future volitional decisions. The agent must be able to volitionally choose the kind of volitional being they will become in the future. This seems like a tall order, but really the seeds of this idea were already discussed in Aristotle's *Nicomachean Ethics*. According to his virtue ethics, a person's decisions and the moral consequences of those decisions flow from the

fact that a person has made past choices that have cultivated them (or not) into being the kind of person they are now, in particular by having reached a point where certain kinds of, ideally virtuous, decisions have been automatized because of habit formation.

To reiterate then, there are at least two kinds of free will demanded by incompatibilists. One is what Kane regards as freedom of action, and the other is what might be regarded as the freedom to choose one's capacity to choose, not right now, but in the perhaps distant future after much cultivation or training. This type-2 incompatibilist free will or metafree will requires the capacity to develop one's nervous system, or the character realized in it, in an intended way. This is the freedom to choose what kind of chooser one will become weeks, months or years down the line.

According to Kane, ultimate responsibility for an action and its immediate consequences requires responsibility for anything that is a sufficient cause or motive of that action. Thus, if an action follows necessarily from having a certain character (which we can use as a shorthand notion for the conglomeration of desires, values, tendencies, capacities and principles that one has), then to be ultimately responsible, even if in part, one must be in part responsible for the character that one has. This hearkens back to Aristotle's insight, in the *Nicomachean Ethics*, that in order to be in part responsible for the wicked acts that follow from having a wicked character, one must be in part responsible for having the wicked character that one has.

Kane recognizes that there would be a logical regress if we chose our present character based on the character we had in the past, which we in turn chose based on the character we had in the more distant pass, and so on, until we came into existence. To escape this regress, Kane argues that there must be a break in the chain of sufficient causes. This happens when a character-forming decision is made that is not sufficiently caused by one's present character or

state. This, Kane argues, requires that this "self-forming" decision is not determined, because if it were, it would be sufficiently caused by the preceding state of affairs. Only under indeterminism is an outcome not sufficiently caused, because identical causes can lead to various different effects or outcomes just by chance. But if a self-forming act or decision were utterly random, then our resulting new character would not have been one that we willed. It would instead be one that just happened by chance. We might make a random self-forming decision, and find that suddenly we are a murderer for no reason at all.

To get around this, Kane in *Four Views on Free Will* (2007) and Balaguer in *Free will as an Open Scientific Question* (2009) zoom in on a very special class of decisions where people are torn between two options, both of which they have willed. These 'torn decisions' might happen only rarely in a person's life, but unless at least one such self-forming act happened in their life, they could not be even partly ultimately responsible for anything they decide or do on the basis of their present character, because of the above regress. To break sufficiency, at least these kinds of decisions have to be undetermined, even if every other act or decision in a person's life is determined.

Kane writes: "…undetermined self-forming actions…occur at those difficult times in life when we are torn between competing visions of what we should do or become….yet the outcome can be willed (and hence rational and voluntary) either way [that we decide] owing to the fact that in such self-formation, [our] prior wills are divided by conflicting motives…When we…decide in such circumstances, and the indeterminate efforts we are making become determinate choices, we make one set of competing reasons or motives prevail over the others then and there by deciding" (2007: 26).

Even though which way we will decide is not determined, since either choice is one we want and have reasons for, Kane argues that it is willed because it reflects our purposes and intentions, and is not utterly random. Indeterminism in essence selects among a class of options, each of which some part of us has agentically specified as one 'we' (note the need for a divided self here) want and intend to do. No matter how we decide by chance, we succeed in doing what at least one part of us knowingly and willingly was trying to do. Note that the reasons and motives for the two options have to be different for this to work, because otherwise we are like Buridan's ass choosing between options that are equivalent to us (see Balaguer 2009: 74-75; it is unlikely that Buridan's ass kinds of decisions will transform our character, so reduce to type-1 libertarian free will choices). Kane considers, as a prototypical example of a torn decision a businesswoman who wants to help someone out of empathy, but who also wants to get to an important meeting, because of ambition, and she cannot do both, because the clock is ticking. She must choose, either/or.

I think Kane and Balaguer do not need to try to rest the possibility of the existence of a type-2 libertarian free will or a metafree will on such rare events as torn decisions, or the notion of a divided self with conflicting sub-agendas or sub-desires, even if this is a valid way of grounding them. All that is required to ground both type-1 and type-2 libertarian free will is that options reflect our reasons and motives, and that the option selected is undetermined.

This is where criterial causation can help. To ground a type-1 libertarian free will it is enough to specify some criterion such as "I need an escape route." Many possibilities can be generated, and one, just by chance, can be selected, whether in our brains or the brain of a tiger. The selected escape route is not utterly random, because it had to be an escape route. But it is not determined, because a different escape route might have been selected by chance. This is what

Kane regards as freedom of action. Kane thinks we need more than this to have true free will, because unless we can intentionally reshape our characters, we are subject to the above regress.

To ground a type-2 libertarian free will, we can imagine, given the constraints imposed by our present character, the kind of future character that we want to achieve. This then sets criteria for the, in part, random fulfillment of those character-defining criteria. For example, in light of our present character, which, let us say, is disgusted by our present lack of integrity in late December, we might set a criterion that we need to resolve to increase our level of integrity somehow. Now, in part just by chance, we might resolve to be more honest, and set that as our New Year's resolution. But we could have just as easily decided, again by chance, to be more kind, or to keep our word better, or to be less impetuous, or less jealous, or less greedy, or more reliable, or to henceforth be a better friend, or be more organized. Let us say, just by chance, the possible resolution to be more punctual arose from unconscious information processing to the level of consciousness for consideration of adequacy. It might then be rejected as not adequately satisfying the criterion of increasing our integrity. Then a new possible resolution might come to the fore of consciousness, perhaps from unconscious processing, say, to keep our word better in the future. Let's say that this passes the threshold, after some conscious consideration, for meeting our integrity-enhancing criteria. So why did our New Year's resolution turn out to be "to be more honest" rather than "to be less selfish and greedy?" Well, just by chance the first one passed our criteria for integrity enhancement first, and became a self-forming resolution. This is not an utterly random resolution, because it had to be an integrity-enhancing resolution. But it is not determined, because it could easily have turned out otherwise, even though we could not change our character at the moment of threshold crossing. However, once a resolution is in place, in part just by chance, it will shape our actions come January, and over several months we may

be able to accomplish the improved character we envisioned. Note that because we chose this resolution for self-improvement in part just by chance, we did not settle our resolution (i.e., determine which resolution we settle on); what we did settle were the criteria that possible self-forming resolutions would have to meet, and an adequate one passed threshold first, which we then went with. Type-2 libertarian free will is a slow and bootstrapping goal-directed sorites-like process because changing one's character takes effort and time.

Criterial causation can ground both type-1 and type-2 libertarian free will, and is not rare, like torn decisions. We are constantly engaging in this kind of criterial decision-making at multiple levels, and also constantly correcting deviations from our paths to our envisioned goals in a cybernetic process involving feedback. This is why both contingency and goal-directed agency are everywhere you look in human action and decision-making.

11. Criterial Causation Overcomes the Luck Argument Against Moral Responsibility

In the previous section, I considered an alternative to the Kane-Balaguer strategy of trying to ground type-2 libertarian free will on a foundation of undetermined torn decisions. Balaguer thinks that it is an open empirical question whether torn decisions are made in the brain in a way that is undetermined. I think my account of neural criterial causation, if correct about how the brain works, would give his or Kane's related theories just the sort of empirical account they need to say that we are libertarian-free.

The Kane-Balaguer strategy of grounding libertarian free will in undetermined torn decisions is vulnerable to the criticism that it fails to overcome the argument from luck according to which, if a critical moral choice goes one way versus another due to randomness (say amplification of quantum fluctuations to neural spike timing variability), then we cannot hold people responsible for the consequences that follow from that choice.

The argument from luck runs as follows: "If decisions or actions occur indeterministically, such that two or more alternative decisions might be made at t, each with a non-zero probability, and everything is exactly the same in world history until t, then there is nothing about the world or the decider prior to t that accounts for one decision being made over the other. Which gets chosen is just a matter of (perhaps weighted) chance, not a matter of agentic influence on specific outcomes. If one decision should turn out better or worse than another, well, that is just a matter of luck and not a matter of agentic choice. But if decisions and consequences are just a matter of luck, then the decider cannot be responsible for the decision made or for its consequences." Note that if the argument from luck works at all, it works not only against libertarians, but against everyone, including compatibilists; see Pereboom (2001, 2014), Levy (2011), and Caruso (2012, 2015).

Note that for Kane and Balaguer the important indeterminacy happens at the moment of choice, not before it. This means that people cannot bias the chance outcome toward the morally superior choice with their wills because even if they could, their decision to bias one way versus the other would itself be subject to the argument from luck.

Neil Levy (2011) in *Hard Luck* argues that free will is ruled out or precluded by luck, regardless of the truth of determinism or indeterminism, making him a hard incompatibilist like Pereboom or Caruso, and a denier of free will. This entails a denial that anyone bears any moral responsibility whatsoever. According to Levy's account, even though Hitler chose to systematically annihilate all Jews he could find, he did so just by luck, whether because of inherited wicked character (constitutive luck, so not his fault), or because of present luck when deciding between options (so again, not his fault). As such, I find Levy's denial of moral responsibility a profoundly nihilistic view of human beings, their choices and life in general.

Those who deny that information can be causal at all, because they accept Kim's exclusion argument, must accept that informational mental events such as willings cannot be causal either. For them, free will and moral responsibility vanish along with the disappearance of mental causation. So free will and moral responsibility vanish for Levy in at least two ways, via luck and via the idea that all causation seeps down to lowest level of physical causation, leaving no room for mental events like willings to make any difference to physical outcomes.

Levy's attack on libertarian free will is rooted in the traditional failure, he says, of libertarians to offer a contrastive account of choices (2001: 43, 90). Consider van Inwagen's (1983) potential thief pondering whether to steal from the church's poor box. He is in a classic torn state before the decision is made. He is torn between the motive to have money, and the motive to honor a deathbed promise he made to his mother to live morally. Levy points out that libertarians fail to offer any explanation concerning why the thief decides to steal rather than not steal. It just happens. And had the decision, by chance, gone the other way, libertarians could not explain why the man refrained from stealing rather than stole. This allows Levy to say that (1) the character we start off with is a matter of (constitutive) luck, so we are not responsible for it or the choices, acts and consequences that follow because of our character that was 'foisted' on us by genetics, the environment and their interaction, and (2) any so-called 'self-forming act' or choice comes down to (present) luck, so we are also not responsible for it or its consequences either. He calls this his "luck pincer." It is really just a variant of the logical regress against the possibility of ultimate responsibility summarized in Section 10. This was of course the regress that drove Kane and Balaguer and other proponents of libertarian free will to rely on torn decisions in the first place. Because of luck, Levy says that torn decisions fail to afford moral responsibility, because all choices come down to constitutive luck, present luck, or both.

Note that in Kane's and Balaguers groundings of libertarian free will in undetermined torn decisions, there is no higher governing basis for making a choice in a torn decision like van Inwagen's thief's; deciding to steal and then stealing the cash just happen. Had he decided not to steal, then not-stealing would have just happened. Levy's point is that if things just happen—and this is almost a Buddhist perspective—there can be no blame. Shit just happens.

The kind of self-forming New Year's resolution I considered in section 10 would not be solely a matter of luck, because it would have to meet the integrity-enhancing criteria set in place by the agent. Even though this resolution might have been chosen versus many possible others, it had to be one that met those criteria, so was not utterly random, so was not solely a matter of blind luck. In contrast with Kane and Balaguer's accounts of libertarian free will, in the case of criterial decision-making, there is a higher, but non-determinative governing basis for making a choice. Yes, that the resolution ended up being "to be more honest this year" rather than "to be less greedy this year" was a matter of luck in the sense that the first proposal passed the threshold for adequate satisfaction of integrity-enhancing criteria first. But it is not an utterly random outcome, like choosing to steal the money or not, as in the thief's torn decision, or choosing to help someone in need or go to the meeting, as in Kane's businesswoman example of a torn decision. Under criterial causation the choice is not utterly random because it had to be an integrity-enhancing resolution. It is also not determined, breaking the chain of sufficient causes that underlies the ultimate responsibility-destroying regress, because a different integrity-enhancing resolution might have won out. The regress is broken by adding indeterminism. But the luck argument is broken by forcing any choice or action to meet criteria set by the agent him- or herself. Kane's and Balaguer's accounts are vulnerable to the luck argument because there is no basis for choosing one self-forming path or another; one set of motives is randomly favored

over the other. One half of our divided self wins over the other just because. On my account, the integrity-enhancing criteria specified by the agent imply that whatever resolution ends up being chosen was willed and not simply a matter of luck because the decision or choice had to meet the agent's self-forming criteria. Because agents play this criterial role in their self-forming decisions, and continually adjust such criteria in a cybernetic or feedback-based process over years of self-formation (am I in fact getting closer to the envisioned future me?) agents are in part responsible for developing negative or positive characters over years of development. Yes there is randomness in terms of which resolution will win, but there is not randomness at the level of the basis for choosing one option over another, as in Kane and Balaguer's undetermined torn decisions.

Concerning type-1 libertarian free will, we are in part responsible for our actions because we set these criteria versus others that we did not set. And we set these versus others because of the kind of agent who we are. We are not completely responsible, because we are not responsible for the particular way those criteria were met (say we chose this escape route versus another), because this was a matter of chance or luck. So we are responsible for choosing an escape route, though not fully responsible for the particularities of choosing this escape route versus others that we might have picked had it not proven adequate first.

Similarly, concerning type-2 libertarian free will or metafree will, we are in part responsible for our characters because we set these criteria for self-forming resolutions versus others that we did not set. And we set these versus others because of the kind of agent who we are. But we are not completely responsible for our characters, because the initial characters or capacities we inherited were a matter of constitutive luck, and the particular way in which the criteria that we set were met (say we chose to be more honest in the new year, rather than less

greedy), was a matter of present luck. So we are responsible for choosing to make a New Year's resolution to improve our character, though not fully responsible for the particularities of choosing this character-forming resolution versus others that we might have picked had it not proven adequate first.

But even if we are only in part responsible for our actions and characters, criterial causation offers both a grounding for libertarian free will of both types 1 and 2, and a degree, therefore, of moral responsibility.

Conclusion

Assuming indeterminism, it is possible to be a physicalist who adheres to a libertarian conception of free will. On this view, mental and brain events really can turn out otherwise, yet are not utterly random. Prior neuronally realized information parameterizes what subsequent neuronally realized informational states will pass presently set physical/informational criteria for firing. This does not mean that we are utterly free to choose what we want to want. Some wants and criteria are innate, such as what smells good or bad. However, given a set of such innate parameters, the brain can generate and play out options, then select an option that adequately meets criteria, or generate further options. This process is closely tied to voluntary attentional manipulation in working memory, more commonly thought of as deliberation or imagination.

Imagination is where the action is in free will. It allows animals not only to consider possible courses of present action (type-1 libertarian free will), but also, at least for the case of humans, it allows us to consider what kinds of choosers we want to strive to become (type-2 libertarian free will). It allows us to imagine learning this language or that, and then, once a choice has been made to become, with effort and practice, a new kind of nervous system that can eventually speak that language. It allows us to lay in bed and imagine flying machines, then go

build one that we have imagined, and thereby change the physical universe forever. It allows us to imagine a better self that we can then set about realizing through practice. And that future self will be able to make new kinds of choices that are not yet open to us now.

References

Balaguer, M. (2010). *Free Will as an Open Scientific Problem*, MIT Press.

Caruso, G. D. (2012). *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Lanham, MD: Lexington Books.

Caruso, G. D. (2015). "Kane is Not Able: A Reply to Vicens' 'Self-Forming Actions and Conflicts of Intention'," *Southwest Philosophy Review* 31, 2.

Dowe, P. (1992). "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory". *Philosophy of Science,* 59, pp. 195-216.

Hume, D. (1739). *A treatise on human nature*. London: Clarendon Press.

Kane, R. (2007). "Libertarianism" chapter 1 in John Martin Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas, *Four Views on Free Will*, Blackwell Publishing.

Kim, J. (1993). The non-reductivist's troubles with mental causation. In J. Heil & A. Mele (Eds.), *Mental causation*. Oxford: Oxford University Press.

Kim, J. (1993). *Supervenience and Mind: Selected Philosophical Essays. Cambridge University Press.*

Kim, J. (1996). *Philosophy o f Mind.* Boulder, CO: Westview Press.

Kim, J. (2005) *Physicalism, or Something Near Enough*, Princeton University Press.

Levy, N.(2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*, Oxford University Press.

Papineau, D. (2009). The Causal Closure of the Physical and Naturalism. *The Oxford Handbook of Philosophy of Mind*. Edited by Ansgar Beckermann, Brian P. McLaughlin, and Sven Walter.

Pearl, J. (2000): *Causality*. New York: Cambridge University Press.

Pereboom, D. (2001). *Living without Free Will*, Cambridge: Cambridge University Press.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*, Oxford: Oxford University Press.

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

van Inwagen, P. (1983). *An Essay on Free Will*. Oxford University Press.

Woodward, J. (2003): *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.