FREE WILL AND DETERMINISM © Dan Bruiger 2019

1. Introduction

The question of free will *versus* determinism is age-old. I will discuss here, instead, how there can be free will *with* causality. Determinism is a concept that does not apply to the human world of reasons and intentions. I will show that it does not apply even to the *natural* world of which human beings are an integral part. On the other hand, the concept of free will applies only to *persons*, who are able to transcend a given state.

In modern terms, the question has become whether the activity of the brain, which is a physical system and a natural organ of the body, fixes human experience and behavior in such a way that the treasured notion of free will is invalidated. I will show that this is not necessarily the case. If one is not in charge of one's own behavior, does that invalidate personal responsibility? Such a question raises the further question, what is this "one"—the self—that is or is not responsible? I will attempt to answer these questions as well.

2. Mind and brain

Mental states seem to follow one upon another in a sequence sometimes called stream of consciousness. What relationship does this sequence bear to the sequence of events in the brain, in the body, and in the physical world at large? It is generally now agreed that the brain is a causal system, meaning that one neuronal state follows from another in time in such a way that the earlier is held to cause the later. We will shortly examine that notion of causality in greater detail. But let us grant it for the moment. It is further widely agreed that these brain states are the cause of what we experience and how we behave. The question then becomes: does the conscious experience of *willing* have any effect on brain states? More generally, is there a "downward" causation (from the mental to the physical) in addition to the "upward" causation in which the physical states of the brain/body/world cause our conscious mental states in particular? Does "mind" affect brain or is it merely an "epiphenomenon" of physical brain states? Does the conscious self have some power or is it "just along for the ride?"[1]

The famous experiments of Benjamin Libet and others make it clear that brain activity precedes conscious awareness in time. Since we accept temporal order as the key aspect of causality, it is clear that conscious awareness or thought cannot in principle be the cause of the brain state associated with it. But that would not prevent it from being the cause of a *subsequent* brain state. We need not be distracted by some metaphysical question about how mental and physical interact. We need only accept that the conscious mental state arising from some physical brain state corresponds also to a brain state that can give rise to future brain states.

The conundrum—that conscious mental states seem to do no causal work—arises from thinking of the brain as a monolithic structure, operating at a single logical level, which is far from the case. Now, I hold that conscious mental states are *representations* of brain states. But representations to whom and for what purpose? The domain of conscious experience is a representation *to* a certain level of organization within the brain. (How this level is

[1] See, for example, Michael S. Gazzaniga *Who's in Charge: Free Will and the Science of the Brain* Harper Collins, 2011 (2009 Gifford Lectures), p140.

physiologically realized—localized or distributed, for example—is a separate question.) The important point here is that it constitutes an *agent* with powers and responsibilities. In common parlance, it constitutes the self. Naturally, for any conscious mental state (such as willing) there is a corresponding brain state of this agent. That the conscious *mental* state (willing) occurs in time after the corresponding *physical* state is irrelevant because the former concerns *future* action. In the *experience* of willing, the agent represents to itself the fact that it is in the brain state called volition. For, the function of this agent includes monitoring its own relation to the organism and world beyond. There is an ongoing feedback loop, which inevitably involves time delay. Consider, for comparison, the time delays in communication between space probes and their earthbound control stations. Or, to use another metaphor, consider the problem of waging war before radio communication. Headquarters gathers reports from posts on the front and returns its orders to them. The fact that events at the front occur before they are reported may hinder, but does not negate, the power of generals to direct the course of the battle based on the reported information. All the actions are performed by soldiers (neurons), regardless of rank or location. These are simply operating at different levels, playing different roles in the command structure, which is a communication structure. It is precisely that sort of structure that eludes a traditional causal account.

3. Cause and intention

Nowadays, cause is usually understood to mean "efficient cause," in Aristotle's sense: the (ideally singular) instrument by which something happens. That might be a thrown rock that breaks a pane of glass (an irreversible action). Or it might be one domino falling upon another or one gear meshing with another, which are actions whose temporal order could be reversed. (The initiating domino could be at either end; the gear chain could run backwards.) Mathematical models are generally reversible in that sense, because t and $-t$ are interchangeable in the equations. Most events in the real world, however, are not reversible.

Though based on observation, mathematical models and the equations that define them are idealized products of definition. As such, they are timeless in principle. The contemplated causal factors are represented by mathematical variables in the model. But in the real situation modelled, there could be an indefinite number of other factors as well. In nature, causes are always multiple; in organisms, they operate in complex feedback loops. Even a single nerve synapse cannot be understood as a simple linear 'A causes B'.

The usual understanding of causality works for modelling and predicting simple physical systems. But it is not the right concept to account for the behavior of organisms. It is not even the right concept to account for the behavior of man-made devices. It is pointless, for example, to describe the operation of an electronic circuit in terms of electrical fields and discharges, of electrons being conducted through wires. A meaningful description must include the *logic* of the circuit, the purposes the device is intended to serve, etc. This type of description is relative to the concepts and goals of human agents who design and use the device. The communication involved takes place among these human agents. But the notion of *intentional description* can be applied to other organisms as well.[2] The principal differences are that humans did not design or create the organism, and the organism itself is an agent with its own goals and internal

[2] See Daniel Dennett *The Intentional Stance* Bradford, 1987

communication. The human observer must put herself imaginatively in the organism's place, to try to grasp the logic of these communications in the context of the organism's own goals.[3]

Like other human artifacts, a model thus conceived through intentional description is a human concept. It is a finite product of specific actions or definitions. That is the best we can do, since we did not make the system ourselves. However, this concept further posits that the real system itself operates on the basis of intentions, although not *human* intentions. The human theorist can most appropriately try to model its operation in intentional terms. But this can mean either (1) *as though the system were an artifact* designed by humans; or else (2) *as though it were a rational agent* that designed itself.

Either way, such a system (like any model of it) is a logical system. Logic does not operate in time. The *if-then* of logic is not the *if-then* of causality. Although a physical system executing a program requires time for the physical counterparts of the logical operations to happen, there is no time involved in logical implication itself. There may be directionality, however. ('A implies B' is often not the same as 'B implies A'.) There may also be circularity, multiple conditions, and the complexity that characterizes organisms. What there is *not* is the possibility to reduce the system to purely causal terms. Nor is there the need to fret over sequence in time.

We are used to thinking of causality in passive mechanistic terms, with no choice involved. Choice is rather the action of an *agent*, which has *reasons* for its actions that have little to do with simple causality, and often much to do with the actions of other agents. There may seem to be no choice involved if we look at the organism on the level of individual neurons, or at the army on the level of foot soldiers taking orders. Historians may give an account of a battle that seems deterministic in hindsight. This does not change the fact that real decisions had to be made in the moment, often on the basis of inadequate information.

4. Determinism and randomness

Causality is a human category, like Kant's categories of space and time. That is to say, it is built into our very perception. It is a subjective and sketchy notion to the extent that people may differ in what they identify as causes, if anything at all. Some people fail to see causes where they ought to, and others see them where they do not exist. Determinism is another matter, for *the only deterministic systems are products of definition.*

The word *determine* is ambiguous. It can mean something an agent does—as in the case of a sailor who determines (estimates) the latitude with a sextant or a jury that determines (decides) the guilt of the defendant and thereby determines (fixes) his fate. Applied to ordinary objects or to scientific objects like atoms, we are in the habit of saying that the cause of an event determines its outcome. However, this sense of the word is a figure of speech that hides behind a metaphysical notion of a power of causality.[4] In truth, however, it is the observer of the event who determines (finds out) what the outcome is. The observer has no power to *fix* the outcome, only a power to *know* it. And this is a power to decide on how to perceive and conceive it. Language, however lends itself to confusing these senses of the word 'determine'.

---

[3] Or, one could say alternatively, nature's goals for it.

[4] Hume famously dismissed the notion of causal power as no more than sequence in time.

We can never know with certainty or entirety the outcomes or causes of physical events. We can only try to determine what they are to the best of our limited ability. This is not the case with products of definition. We can know with unassailable certainty the results of arithmetic operations because *we have defined them,* along with the concept of number. They seem to be *a priori* truths about the ways we can count, rearrange, and group individual objects in the real world. But this is because arithmetic arose as a generalization of such universal properties in the first place. We can test the generalization as often as we like and it always seems to hold. (One apple plus one orange equals two pieces of fruit.) Arithmetic is more than an empirical generalization, however. For, what we actually then do to entrench such confidence is to create a distinct conceptual system in which these operations are *true by definition.* ('One' plus 'one' is *defined* to equal 'two'.) Such a conceptual system is a *deductive system.* Arithmetic, geometry, logic, equations, and mathematical models are deductive systems in this sense. So, in fact, are machines and all human artifacts, insofar as they are products of definition. Of course, they may also be physical objects with their own reality in addition to their ideal definitions. The real machine can wear out or break down, the vase can be broken. But the machine or vase *as a conceptual ideal* is no more subject to time or degradation than the rules of arithmetic.

Determinism is a property of deductive systems, not of real systems. That is because we can have certain knowledge only of deductive systems—the things that are what they are because we define them to be that way. We can determine (know) their properties because we determine (produce) them in the first place. On the other hand, unless they happen also to be agents, natural things have no power to determine anything whatever. That is because the relationship of determinism is logical implication, not causal or physical power. Yet, our best effort (and natural inclination) is to substitute deductive systems for real systems because the former can be perfectly known. That is what scientific modelling is about. It has proven an enormously successful tactic in predicting the behavior of simple systems that correspond well to products of definitions—namely, linear equations. But this does not make it reliable in principle.

Scientific models are deterministic, but it makes no sense to think that nature itself is (or isn't) deterministic. Similarly, it makes no sense to think that nature is (or isn't) intrinsically random. Determinism and randomness (indeterminacy) are notions concerning our state of knowledge, not the state of the world. To say that an event is random is no more than to say that we cannot determine a cause for it. There are physical reasons for this inability. There are limits to observation and accuracy of measurement at the human scale, as well as at the micro and cosmic scales.[5] If there are limits to what we can know about deductive systems, they are not physical.[6]

To return to the brain, we can model it either as a physical system or as a logical system.[7] Either kind of analysis is a human construct, not to be confused with the real system itself. Yet, we can (and must) try to fathom its logic as well as its causality. The brain is a network of

---

[5] At the human scale we have been spoiled by interactions with the world that have only linear effects; "chaos" is unpredictability owing to non-linear effects. In the micro realm, the system observed and the medium by which we gather knowledge are of a similar scale, so that the act of observation or measurement itself produces physical effects that introduce uncertainty. The extremely small and the extremely large (but distant) are both at the limits of observational capacity, and measurements in both domains are essentially statistical and depend on long chains of assumptions.

[6] See section 5 below.

[7] With the proviso that the logic involved may not correspond to that conceived by human beings.

connections, which exist for reasons as well as from causes. As scientific observers, we are brains that try to understand themselves as such a network.

5. What is the self?

Free will requires the ability of an agent to act autonomously—that is, without its action or experience being unilaterally fixed by events in the world, including events in the brain. The agent in question is a brain function. In the case of human beings, at least, this agent is known as the self.

Let us return to the fundamental principle that all experience and behavior is co-determined by the organism and by its environment. While the outside observer is at liberty to emphasize one input or the other, organism and environment always operate conjointly and interactively. This means that the organism has by definition a degree of autonomy in relation to the world surrounding it. An organism does not merely react like inert matter but *acts* on its own initiative and with its own energy, as part of its attempt to maintain itself as a distinct region of the world. But, what about the autonomy of this particular brain function in regard to the rest of the brain and organism?

I propose that the nature of consciousness is a virtual reality created by the brain function we know as the self. The job of this function is to monitor the state of the body and its environment, and the constantly changing relationship between them, in order to facilitate central control. The bodily self-image is then an "avatar," a character within this virtual reality. Because the conscious self is a brain function (just as literal virtual reality is a computer function), consciousness depends on the brain and its internal communications, just as the virtual reality depends on a physical computer. As already noted, this can mean a time delay between different parts or functions within the brain. While the processing time for conscious perception and volition is longer than that for reflexes and other unconscious processes, the key difference lies in the different roles they play.

For example, there is a reflex such that the hand automatically jerks away from contact with a hot stove. If that reflex is not quick enough to avoid tissue damage, there will follow an experience of pain. The latter involves more complex processing, hence the time delay. But it also serves a different purpose: to insure that the behavior of the organism does not lead to *further* tissue damage. The goal is not to avoid the original stimulus but to protect the wounded tissue during healing. For this reason the pain may persist over some time, if only as a diminishing sensitivity. This goal cannot be accomplished by reflex or automatism, for the same reason that the overall conduct of the creature cannot: dedicated programs cannot cover all contingencies. The extra processing involved in consciousness[8] is called into play when simple automatisms are inadequate. This requires more brain cells which means bigger brains. Smaller creatures compensate for lack of flexibility with greater numbers to maintain species success. Larger creatures put their eggs in the intelligence basket. An insect does not need to be conscious or have a self.[9]

---

[8] I do not mean self-awareness but sentience that is registered as sensation or other phenomenal content.

[9] Insects do not generally demonstrate protective behavior toward damage, such as loss of a limb, but seem simply to try to carry on despite the disability—suggesting that they do not feel pain.

6. Self-transcendence

Granted that the self is a brain function, neither it nor consciousness can exist apart from the body. Its specific role as trouble shooter exists because of the need to transcend fixed programs, such as automatisms and reflexes, which may be inadequate especially in novel circumstances. Genetic adaptation is slow, occurring between generations. It works well enough for simpler creatures with large numbers of offspring. Larger creatures, with fewer offspring, must be able to adapt individually as well. The key to such adaptation is the ability to self-modify in real time. The monitoring function of consciousness would be pointless if it did not serve an adaptable motor system. Reflexes are hard-wired motor responses to specific inputs, whereas monitoring implies choice of how to respond to changing or novel inputs.

A nervous system is deterministic only if it corresponds to a fixed deductive system.[10] A reflex or automatism might be characterized as such a system, which might be written as a computer program. But what to say of adaptable systems? A *given* state of an adaptable system might correspond to a fixed system. Yet, the system can transcend that state by substituting for it a new state corresponding to an expanded system. That too might be characterized as a deductive system, but the process can be repeated ad infinitum.

The very existence of a monitoring system implies the organism's adaptability. However, there is a secondary way that such a system can self-transcend. That is through explicit self-reference. An adaptable system models the world in order to monitor it. If it also models itself as an element of its representation of the world, it can model its own interactions with the world. This means it can potentially understand its own participation in co-creating its situation in the world—its role in shaping its behavior and experience. That is the function of self-consciousness, which is the awareness of subjectivity. It is inseparable from an awareness of responsibility.

6. Responsibility

From the point of view of an observer, free (non-deterministic) behavior and randomness have in common that both are unpredictable. But an *agent* is held accountable to other agents for its choices, whereas random processes are not. In fact, the latter are appealed to precisely when one wishes to avoid responsibility for choice (as in tossing a coin).

A system with a finite number of fixed elements cannot be considered an agent in this sense. On the other hand, a system that can modify itself is a moving target. It is not predictable in the way that a fixed system is. If it acts on its own behalf and is responsive rather than reactive, it may be considered an agent, however predictable or unpredictable.

The ability to predict the behavior of agents relies in part on past observation and in part on the commonality between agents. While one may or may not be able to know the physical *causes* of a system's behavior when regarding it as a deterministic system, one may be able to guess or anticipate the *reasons* for the behavior when regarding it as an agent whose considerations resemble one's own. An agent that is recognized as a human *person* is thus held

[10] At least *some* deductive systems are able to transcend their own structure. This much we owe to Gödel.

accountable to other persons according to the commonly accepted rules and mores of society. One can is thus responsible for events that one is thought to cause, especially if they are deemed to be consciously intended.

Appeal is often made to the deterministic viewpoint to excuse unacceptable behavior ("my genes made me do it!"). This appeal asks us to consider the person to be a causal system rather than an agent with reasons and intentions. This is somewhat a contradiction in terms, if we hold such an agent to be *capable* of transcending a specific configuration—in other words, to be at least potentially non-deterministic. An agent from whom we expect such capability is then *responsible* to reconfigure itself in such a way as to transcend the apparently deterministic aspects of behavior in question. An agent with this capability is deemed to have free will and also the attendant responsibility.

The question of free will versus determinism is not a metaphysical question, about which "principle" reigns in the universe. It is rather the question of whether, in a given instance, a particular agent has effective free with, with attendant responsibility, or should instead be considered a deterministic system (and hence not an agent at all). Ultimately, it is a question of expectation. This is not a matter of which among causal factors is "responsible" for an outcome, but whether causality or intention reigns in that situation for that agent. Causes do not involve responsibility in the moral or interpersonal sense. Only persons are responsible, because they are both agents with their own intentions and members of a group with common rules and expectations. One cannot have the proverbial cake and eat it too. If we cherish the dignity of free will, then we must the accept the responsibility to others that is built into it. Either one is a thing with no will and no responsibility or one is a person with responsibility.