

Retaining Free Will in a Deterministic Machine

Charles E. Pippin

CS 7695 – Philosophy of Cognition

Professor: Nancy Nersessian

December 6, 2003

“I knew you were going to do that, it is because you are predictable”, Thomas is fond of saying to me. He only thinks I am predictable, and in any case he does not understand that I am going to do as I do whether he knows the outcome or not. You see, Thomas is a firm believer in causality. Having been friends for some time, it naturally follows that Thomas usually knows what I am going to say before I say it. So much in fact, he even thinks that I am so set in my ways that I have little choice but to do what I always do or he might even go so far as to say that I am a slave to my actions.

But here is where Thomas goes astray. He believes that because he think he knows what I might do or rather what I likely will do in a given situation, that this implies that is what I must do. If our actions are predictable, then do we really choose them? This specious notion of predictability is often intertwined with causality by him and the like.

Predictability is a tall order, however; and we will soon see that causality and predictability are independent properties, meaning I have the freedom to choose as I wish.

If you will allow, before we address this notion of predictability, let’s visit causality in more detail. Thomas would say that everything we do, the sum of our actions are performed for a reason. They are *caused by* something. For instance, I got a drink of water this morning because I was thirsty, and I was thirsty because some sensor in my brain detected low levels of water in my body and triggered a feeling of thirst. This low level of water in my system was *caused by* the fact that humans need to imbibe every so

often. This works well as an example of behavior being caused, but what causes me to perform more complex actions?

“Of course,” Thomas says, “every cause produces an effect, and so if I know the causes, I will know the effect.” You can’t really blame Thomas, his point is borrowed from Spinoza, “If anything is a cause, the effect as arising out of the cause will be deducible from the idea of the cause, and arise out of it. *Ordo et connexio idearum idem est ac ordo et connexio rerum* (Smith, 144.)” Their argument is a seductive one; after all if we can believe that thoughts are caused, then actions are caused, and as such cause and effect are related. Therefore if we know the cause, we can logically derive the rest.

Randomness

We need to press a little more on this point about free choice, if causes and their effects are necessarily related, then does this imply that a given effect must follow a given cause? This would certainly repudiate the notion of freedom in choice. There is an alternative, Thomas, and that is randomness. Consider that our thoughts are not caused at all. They are just that, our thoughts. It might be considered easier to have freedom of thought, if our thoughts are not constrained. Must everything be caused? Why would I choose to take up piano lessons for instance? There is no mechanism in the brain that detects low levels of musical culture in my system and triggers a piano playing thirst. Yet, one morning, recently, I awoke and decided to inquire about piano lessons. There were no piano lesson lobbyists within the confines of my skull. What caused this idea, then? It is easier to believe that “I just decided to do it, nothing caused it.” It was freely

chosen. This randomness in choice, libertarianism, suffers, however, from a grounding problem (Churchland, 2003.) My ideas had to come from somewhere. As an alternative, if we consider then that nothing caused my choices at all and that they are purely random, then I, as an agent am purely random and therefore not really free to do anything at all. I am a slave to randomness. Along these lines, if a random thought enters my mind to take Ballet lessons, what is there to save me? To be fair, we don't as of yet understand the inner workings of the mind. There are too many variables, at the neural level, events may be quantum and indeterminate in nature. However, if we consider that at the neural levels, thought could be a set of random events, then what is left to define me as an individual when all of my actions, rather than being based on causal events, are based on indeterminate chaos? Hume would argue, and I think Thomas would agree, is that our actions are caused by events occurring in our minds. That is, the idea that our thoughts cause our actions, and our thoughts are affected by our beliefs, past history, the environment, and our individual self (Churchland, 2003.)

"Isn't that what I said?" Thomas relates. Unfortunately, we are not finished yet; consider that if everything I do is *caused by* something, then how can I be free to choose, when really all I am doing is following a deterministic causal script. This puts us in a bind, because we either have our actions resulting from a set of random events, and therefore we cannot be held accountable (or free) or they are determined by a causal script, and we therefore cannot be held accountable, or (Pinker, 178.) We are not convinced, however, by this deterministic view of choice, the idea of each individual acting according to what their nature deems necessary (Smith, 141). It commands that, when we are hungry, we

seek food, if we are lonely, we seek companionship, and if we are angry we seek revenge. According to this view, rather than the mind leading the body, the body leads the mind because for each cause we have a related effect and our actions are tied to that effect. In such a hierarchy, Spinoza would say to blame an individual for their actions would be as futile as blaming a triangle for not being a circle. What Thomas and I want to know then is, how can the body lead the mind? Or to put it another way, has any body ever designed a work of art or written letters to a loved one without the mind to guide it? A true behaviorist, Spinoza replies,

“I have shown that [those who make this objection] do not know what the body can do, nor what can be deduced from its nature alone, and that they find many things are done merely by the laws of nature which they never would have believed possible without the direction of the mind. (Smith 150)”

Given that direction, a mind follows the needs of the body. The body *causes* the behavior of the mind, because cause and effect are related. That is, given a cause, here a need or desire, then we know there is an associated effect which the mind follows.

Ah, but Thomas, don't forget about time! Hume reminds us that causality necessarily includes a time factor. A cause happens *before* an effect. One cause may produce multiple effects, and Hume points out that cause and effect have an implied order; determining ahead of time which effect will occur from a given cause would be similar to an assertion that all men are married because all husbands are men. (Smith, 225)

Therefore, cause and effect are still related, but this relation *hinges on time*.

As a result, we are not caused to perform action y because x happened, but rather we are caused to choose what to do with Y when confronted with X. To put it another way, for

the self to be held accountable for behavior and to behave in a consistent manner, it must understand the outcomes of choice. The mind, moreover, must consider how actions and effects will affect the self. If there was no causal relationship between cause and effect then, the self could never be held accountable for their actions, one could never improve one's behavior.

The piano lessons remain on Wednesday evenings, and it is something I now feel obligated to do. Afterall, I have paid money, invested time, and purchased a piano. Nevertheless, I am still free to choose to go to lessons, or to stay home and watch Jerry Seinfeld on TV. It is true that my earlier choice caused my future choices on Wednesday evening to become more restricted, but I have choice nonetheless.

Predictability

While we can agree that every event has a cause, it is more difficult to derive the full set of causes. If such a set of causes exist, it does not imply that we know them and related conditions and therefore we cannot predict the precise nature of an event. Consider the set of brain states and conditions that may exist to cause thoughts which cause our actions. Knowing that this causal structure exists does not grant causal knowledge. predictability. Without predictability, however, we can still have causality, we are just unable to predict outcomes (Churchland, 205.) Causality and unpredictability can co-exist then and we can retain free will. So far, so good. Ah, but what if we did have predictability? Let's say that Thomas really can predict what I will do in a given situation, will I comply? That is to ask, if outcomes are known a priori, are we free to

affect them? Of course not, Thomas would argue: Given a set of causes and conditions you will always do exactly what you are supposed. You have no choice but the choice as determined by the set of causes and conditions, you only have illusory choice; your outcome will follow the path of a set of causes, given conditions, and predicted effects.

What Thomas is asserting then is that given causality and predictability we lose free will and therefore accountability. I could not be held accountable for my actions if it was known in advance that I was going to perform them. I think Thomas may actually belong to that fatalist camp who believes that I cannot do that other than what it is I will do. So here we have another fork, if my actions are not predictable, I have free will, and should be held accountable; but if they are predictable, I have only freedom to perform that set of actions and therefore cannot be held accountable. Free will then becomes a function of an outside observer; only if this observer is performing predictions of my causality then I *must* do what I will do. Otherwise I *can* do what I will do.

This notion of predictability is a bit much, even for Thomas. That is, unless he has a direct line to God. Predictability of behavior is, after all, akin to omniscience.

Probability theory provides a more tenable solution. Consider that while we are unable to observe all conditions and causes of an event, we can observe the higher level, more salient set. We can then observe behavior for a given period and make predictions based on likelihood of occurrence. With some accuracy, we might then have a suitable probability of outcomes prediction machine. But there is a problem with this approach, even if I can predict 99% of the time that I will not race my car through the streets of

downtown Atlanta, does that mean that I will never cause pedestrian distress? The probability machine can only predict likelihood, it is not a true predictive model and therefore you do not know what I will choose; only what I will likely choose. To make this even more difficult, these probabilities would be very difficult to calculate, if we could even begin to calculate them, this is because there are so many factors that can affect our behavior.

Drugs, Hormones, and Feeling

Before we consider further the power of choice, and how to predict it, we should visit the limiting power of choice factors. As human beings, we are easily enslaved to factors of the flesh. If we are hungry, sleep deprived or feeling very emotional our decision processes can be affected. My brother is more likely to choose a grumpy outlook if he has not eaten in a few hours, and I am likely to alter my behavior if I haven't slept well the night before. Further, from the age of puberty, increased sexual interest is induced by a flood of hormonal changes on the brain; this for many causes us to alter our behaviors in goals in given situations.

In a similar vein, pharmacological interventions such as Prozac, Ritalin, Lithium, and others are commonly used to change our desires, and alter our behavior. While some of these are natural affectors, all can have a powerful effect on our decision architecture and processes. A recent trend is the use of drugs to treat behavior disorders in children. In just a few examples depression is at times treated with Lexapro, and attention deficit with Adderall (Kluger, 50). These drugs alter a brain state and make it easier for an individual

to choose correct a behavior. We might ask how can we be free to make our own decisions if our decision architecture is so malleable? We must then consider all inputs to our decision process, and not just relevant events when we consider the causal nature of our decision process. This flies in the face of dualism, proponents of which might argue that our decision substrate is a constant; our souls effect our actions independent of the brain. It is evident however, that our decision architectures are easily affected by our feelings, relative health and hormonal variance.

The Choice Machine and Self

Damasio presents this notion of the decision architecture as being guided by the individual, with a requirement for individuality having a consistent knowledge of self. This is what allows for consistent, behavior, relevant to an individual, rather than random instability. He presents two necessary representations for knowledge of self, the first being “key representations of events in an individuals autobiography.” These are what define us a person, and secondly, “primordial representations” of an individual’s body (Damasio, 239.) That is, in order to define the self, we must have as inputs the set of conditions that define our history, and the set of environmental conditions (emotions, feelings, perceptions) that define us lately. The self represents an overarching process that considers outcomes of different choices, given the information known, and selects a choice that maximizes utility.

This presents human beings in the light of “choice machines, to borrow a phrase from Gary Drescher and explain it according to Daniel Dennet: Choice machines take a look at the world and evaluate their set of options based on what they believe a likely outcome might be. A choice machine can consider what would happen if they did A versus B, and then make a free choice as to whichever path provides the most value (Bailey, 27).

Our freedom to choose then, is a product of not only cause and effect but also of weighted condition factors. If we can divide our existence into a set of discrete time slices, with each slice representing a single decision, then for that decision space we have a set of known conditions, or environmental input, known history conditions (past brain states), and the current brain state as input into a decision mechanism.

Reconsidering this probability approach to predictability, there are many possible outcomes for each situation, the best you can do is a likely guess. For instance if President Kennedy had had access to this machine during the Cuban missile crisis it would have only given a reasonable estimation that the Soviets would back down, he still would not have known for sure how a given outcome would have affected millions of Americans through the choice of his adversary. No, he would have a more refined general predictability , but he would not have had precise predictability. To this, Thomas agrees. “Yes, a probability machine is not sufficient. What we need then is a sufficient prediction machine.”

A Sufficient Prediction Machine

To see where this will take us in our discussion of free will, let's entertain Thomas in his experiment. We consider a decision process for a finite time slice to take as inputs a set of perceptions, the current state of self and knowledge of historical states of self, and the output being a set of actions that will serve to maximize an agent's utility. As part of this, this process is based on the laws of cause and effect, such that given a set of events and conditions, X, they will generate a set of effects, Y. Further, given the exact same set of conditions we will always have the exact resulting set of effects. The idea then, is this: if we can build a model of the decision process in a machine, providing it the correct inputs at a given time slice, it will predict the exact outcome of that process. If we are able to provide it with the same inputs that I am receiving when, for instance, I consider whether to attend piano lessons, we will be able to predict whether I will choose to attend them. Effectively, this is a notion of modeling the mind's decision architecture as a set of equations, albeit very complex ones, that can be implemented in a machine and solved.

This idea hinges on the knowledge of the current conditions. Such knowledge is difficult at best to capture. Consider an example from Churchland: if we drop a dollar bill from the Eiffel Tower, we know that it will fall to the ground, but we do not know the exact trajectory of its path, nor where it will land. Subtle wind currents, causing force variations on different points of the dollar bill, will contribute to its fluttering pattern. We know that these conditions cause such movements, but we cannot take the measurements fast enough nor compute the outcome fast enough to predict these patterns (Churchland 205.) We can consider other classical systems such as weather patterns or

even the beginning of a simple game of pool, in which we only have one ball striking an arrangement of fifteen others. It should be simple to predict where each ball will stop, given the laws of physics, but we cannot capture the subtle changes in velocity, striking angles, friction and so forth, much less provide computation that would perform the prediction calculations in real time. One might consider such precise predictions to be impossible.

Consider further the set of inputs that our brains must receive at any given time slice and then process in order to make a decision. How might these even be captured, and represented in a meaningful format? As I stand on the curb debating whether to cross the street, I can see cars approaching, and hear the hum of their engines against the road. The problem is, even if you are standing right next to me, you may not see the same thing, there might be a telephone pole or another person blocking your view. If you were holding the prediction machine, this would introduce subtle error into the inputs and this error would be magnified against other errors and in the prediction calculations, those calculations would not be precise. Even if you could see the cars from the same angle, how would you represent it so that it can be usefully transformed in regards to other inputs, such as the sound of the cars, or the desirable smells emanating from the coffee shop across the street. Further, other inputs to my decision to cross the street would be knowledge of past decision states. I might have never been to the city and prefer to be cautious or I may have just witnessed a vehicular homicide within the past hour and have no desire to step off the curb.

Thomas would resolve the representational issue with grounding principle. His idea here is rather than capture modal representations of the environment outside of our skulls, instead, simply capture exact molecular snapshots of the neural substrate for a given time-slice, and then run simulations to predict the next time slice. That is, if we are not in the dualist camp and can assume that our selves are implemented within the confines of our skulls; that set of neurons and synaptic connections that implements our current brain state and are able to represent all past brain states; then if we were able to take a snapshot of our brains at a give time slice, we would have all of the inputs to the decision process. Moreover, no one has yet shown that at the neuronal level, events happen without cause, so for our purpose, this substrate does appear to be a causal machine (Churchland, 204). Given then this snapshot, we would have the exact set of events and conditions, X that will generate a given set of effects, Y.

Thus far, we will have solved the representational issue, but the issue of capture now looms larger than before. Thomas's answer here suffers from futurology, relying on continued progress in brain imaging. "Think of an fMRI on steroids," he interjects. With current technology for safely measuring brain activity in humans, we can get an idea of regional changes in brain activity over time, unfortunately the best techniques that we have now, fMRI and positron emission topography, only get us to a spatial resolution of about 2 mm, where a cubic mm will contain about 100,000 neurons (Churchland, 18.) In addition, these readings can take a number of seconds, while our requirements would dictate nearly instant snapshots. Granted, such technology is not unimaginable. Following this idea, if we can posit that brain imaging technology may advance to the

point where we can take exact neural snapshots of the human brain, that is, to capture the exact electrical and chemical pattern of the brain, at a given time slice (in real time), down to the level of single neurons, and for every single neuron, then it will be possible to capture our set of input conditions, X.

After assuming a suitable capture mechanism we are still faced with the complexity of performing relevant computations, and performing them in real time. Consider the enormous complexity of the brain; it is a system that, by conservative estimates, has about 10^{12} neurons and 10^{15} synapses connecting these neurons (Churchland, 205.) Then if we define a time slice to be relevant to the scale of neuronal events, in the millisecond range, then to model a time slice of the brain, we must model about 10^{15} parameters! This problem is computationally intractable. Even if we had a machine large enough to represent all of these variables (theoretically possible since this number is on the order of the number of grains of sand on a beach), iterating through every possible combination could take longer than our lifetime. Further, even if we could employ strong heuristics in the search through these combinations, the calculations would still take much too long to predict an event in time.

Another possibility then is to use a quantum machine. A quantum machine could, again in theory, perform every possible combination of events instantaneously and provide us with an answer, before the real outcome, thus, a suitable prediction. Here again, however, we are reliant on futurology. “Don’t forget, the brain is able to perform these calculations”, Thomas reminds us. The complexity of the model need only be as

complex as what it is modeling. So here, we have an implementation already to follow. Perhaps a better model then, would be one in which we model the brain exactly in our machine. A simulator, “in a vat”. Each neuron in the brain is a neuron in our machine, and each synapse in the brain has an equivalent as well. At any given time slice, all we need do is set each of the condition variables in our brain simulator and allow the simulator to “run” for an additional slice, as the real brain would in the world. This would provide us with a suitable prediction machine, theoretically in real time.

As alluded to earlier, what Thomas is describing smacks strongly of religious overtones. After all, given a machine that can predict exactly what an individual will do, and what choices they will make, isn't that akin to an Omniscient God who knows all of our actions in advance? Thomas reply is “Not exactly, we are proposing only a single time slice prediction machine.” What he means by this is, we are not creating God. That would be the same as creating a machine that is Omnipotent, Omniscient and Omnibenevolent. He is simply proposing the plausibility of a machine that can predict single time slices of behavior for a given individual, given the full set of captured input events. Fair enough, Thomas, we can grant this thought experiment for the purposes of our discussion. Let us assume, then, that such a machine is in Thomas's possession and that it was available for him to use in predicting an individual's behavior and that that individual is me.

Fatalism vs. Determinism

What Thomas would like us to believe is then, that our behavior is controlled by this machine. The idea is that given the prediction of the machine, we cannot perform any other act than the act we will perform. For example, if I am debating whether to attend my piano lesson on this evening, Thomas might use the machine to predict what I will decide to do, in the exact instance that I am deciding. Here, the machine is given the inputs of what I know about past piano lessons, my desires, how I feel that day and my knowledge as to which *Seinfeld* episode will be shown on TV during my lesson. In effect, the machine has the same exact information that I do in regards to this decision, because the machine has captured a snapshot of my neural state.

Imagine then that this machine, much to my dismay, predicts, despite my capricious nature, that I will elect to attend my piano lesson, rather than stay home and watch TV. It is as much as saying that given conditions the way they are, this is what is “going to happen.” But, let us first consider an example from Daniel Dennet, in his discussion of human choice machines:

“Going to happen” is a very misleading phrase. Say somebody throws a baseball at your head and you see it. That baseball was “going to” hit you until you saw it and ducked, and then it didn’t hit you even though it was “going to...” People confuse determinism with fatalism. They’re two completely different notions.

In a sense, the fatalist view would agree that if an event were predicted to occur, that it must occur, and there is nothing I can do about it. I would in fact be prevented from choosing any other choice and would only be following a predicted trajectory. What if I were to “duck” this ball then and choose as I please? According to Thomas, “That is exactly what is predicted. The machine predicted your choice.” My free choice then is

an illusion, as he would have it. While we might both agree that I do not have to follow a set trajectory, the fatalistic dogma, but can choose which path I might follow; his supposition is that the machine will predict that path. It will predict in advance my dodge of the baseball, so to speak and yet still determine my choice. This is where our notion of free choice is hinged: if my choice is determined is it a free choice? Thomas doubts that this is the case, let us take a deeper look.

In Thomas's ideas about choice, the freedom to choose is built upon the complexity of the factors involved in the choice, the sheer number of possibilities. Solve the complexity, and you remove the freedom from the choice. However this distracts us from the real meaning of freedom of choice, if we consider the notion of what it means to be free, we are truly free to perform an act if nothing *restricts* us or prevents us from that action.

We seek support for our argument from Susan Wolf in regards to performing an action X: if it has been predicted that we do X, then it "must be the case that something prevents one from doing anything other than X or perhaps that something interferes with one's ability not to do X. (Wolf, 115.) But that is not the case at all. In fact, nothing prevents me from exercising my knowledge, skills and abilities to decide my actions. I am in fact free to make a choice because given a situation I *could* choose otherwise. Knowing what I *will* choose does not in any way limit my ability *to* choose.

Conclusion

As a result, even if Thomas could precisely predict my actions, he would only be reviewing predictions of what I will *choose*. It is no different in this sense to look back on what I did choose than it is to look forward on what I will choose. The choice is still mine to make and a prediction machine does not force me to make one choice or another and does not in any way prevent my choice. It only reviews my choice, albeit, in advance. It can be thought of along the same lines of a doctor telling an expectant mother the sex of her baby, through the use of sonogram technology, before the actual birth. Telling the mother that she will have a boy does not force her to have a boy, it only provides information in advance of an outcome that would otherwise not be known until later in the future. While we could argue that with a conception, the outcome had already been formed prior to the sonogram or birth, but was not yet known, the same argument might hold for my decision architecture, when we consider it as a processor of a set of conditions and effects, X. Of course, given the same set of inputs to a mathematical equation, we will always get the same outputs, whether we run that equation now, or in advance, in a prediction machine, or within my skull. Even if this suitable prediction machine exists, and Thomas were to possess it, it is of no consequence to me that my outcome may be known in advance, because in his knowing it, it is still in no way restricted. My choice remains mine own. As such, we can see that free will and predictability (as well as unpredictability) are no less consistent: free will being based on causality, I remain free to select my actions, whether Thomas can predict them or not. Afterall, Thomas, one might think you would have known I was going to say that.

References

- Bailey, Ron. (2003.) Pulling our Own Strings: Philosopher Daniel Dennett on determinism, human “choice machines,” and how evolution generates free will. *Reason*, May 2003, 25-31.
- Churchland, Patricia Smith. (2002.) *Brain-Wise: Studies in Neurophilosophy*. Cambridge, Massachusetts. The MIT Press.
- Damasio, Antonio R. (1994.) *Descartes' Error: Emotion, Reason, and the Human Brain*. New York. Penguin.
- Kluger, Jeffrey (2003.) Medicating Young Minds. *Time*, November 3, 2003, 48-58
- Pinker, Steven. (2002.) *The Blank Slate: The Modern Denial of Human Nature*. New York. Viking.
- Smith, Norman. (1902). *Studies in the Cartesian Philosophy*. London: Macmillan and Co., Ltd.
- Wolf, Susan. (1990.) *Freedom Within Reason*. New York: Oxford University Press.

Footnotes

- 1) This estimate is commonly known, but an explanation can be found on:
About Big Numbers. (author unknown)
<http://pages.prodigy.net/jhonig/bignum/>