1-2016

More Wobble: Daniel Dennett's Elbow Room Unscrambled

Patsy Cadareanu
pxc7985@rit.edu

Recommended Citation

Cadareanu, Patsy, "More Wobble: Daniel Dennett's Elbow Room Unscrambled" (2016). Thesis. Rochester Institute of Technology. Accessed from

Patsy Cadareanu
Dr. David B. Suits
Philosophy Senior Thesis
December 7th, 2015

More Wobble[1]: Daniel Dennett's *Elbow Room* Unscrambled

Through the use of the seemingly endless metaphors in his books *Elbow Room* (1984), *Brainstorms* (1978), and *Freedom Evolves* (2002), Daniel Dennett is able to construct a coherent stance on what it means to be morally responsible in a world that may or may not be determined. This paper explores Dennett's notion of responsibility and its implications.

Determinism is the doctrine that all events are necessitated by previous events which have caused the only possible future. In contrast, indeterminism is the idea that events do not come about through necessity. Indeterminism is often thought of as 'freedom,' but this is not so for Dennett who, like Hume[2] before him, is a soft determinist (or compatibilist) and espouses the notion that deliberation and the free will associated with it is compatible with determinism; in other words, Dennett believes that knowing that a system is deterministic does not reveal anything about the necessity of the system. The free will he refers to is not the absolute freedom of one like Jean-Paul Sartre, but instead consists of what Dennett refers to in the subtitle of his book *Elbow Room* as the 'varieties of free will worth wanting'; these relevant freedoms are self-control and appropriate, meaningful deliberation.

Dennett's idea of responsibility relies heavily on the structure and necessity of the process of deliberation. Thus, to fully understand and evaluate responsibility, the strength of Dennett's deliberation technique must be tested. This will be done methodically: **first**, control will be defined and explained as the root of Dennett's argument through the exhibition of its relation to determinism. **Second**, the established definition of control will be used to define a self as a locus of self-control; the relationship between skill and luck will also be introduced in this section, and its relevancy explained. **Third**, Dennett's process of deliberation, especially as it

---

[1] Robert Nozick found that 'more wobble' is almost an anagram for 'elbow room.'
[2] See "Of liberty and necessity," in *A Treatise of Human Nature* (2.3.1–2).

coincides with determinism, will be explained; the difference between local fatalism and determinism will be discussed and *real* opportunities will be explained in terms of the luck discussed in the previous section. **Fourth**, the principle of "could have done otherwise" will be explained as a foundation for judging moral responsibility and, using the ideas from the previous parts, Dennett's argument against this principle and his suggestion of a replacement principle will be shown. **Fifth**, the intentional stance will be introduced and its effectiveness judged; the use of the stance in deliberation will also be considered. **Sixth**, through the analysis provided in the previous parts the reader will have achieved a wholesome understanding of Dennett's compatibilist system and see how these ideas are applied to create a new idea of responsibility. Dennett's "arbitrary responsibility" will be compared to Sartre's absolute responsibility so as to provide a contrast with a philosophy where determinism and choice are not compatible. **Finally**, a quick criticism of Dennett's philosophy will be noted.

Part I: Control as the Heart (that is, pumping mechanism) of Dennett's Thesis

Control requires at least two agents: the controller, call it A, must have both a normal range of states and desires (or something "like" desires; these can be programmed into A, for example), and the controlled, call it B, only requires states but can also have desires. For A to control B she must drive B into one of B's normal range states which A *wants* B to be in; further requirements for control include that A must stay in contact with B, and A must understand B's states by identifying the parameters of B's operation to control him. In a case where two agents (though B does not have to be an agent) are attempting to control each other (or to resist the attempt to be controlled by the other), there could result a competition for information because whoever knows more about the other's states has more control over the other.

The past cannot control a person: "causal links are not enough for control" because control requires an agent A with information on B's states and a desire to change them, and the past does not hold any information about the present (*Elbow Room* 72), and in any case, the past is not itself an agent. Similarly to time, the environment cannot control a person because it is not an agent: recall that an agent which can control something must have desires--such as the desire to control the other--and evolutionary processes do not operate so as to fulfill desires. Through its cues and one's responses to these, however, the environment has produced people to be the self-controllers that they are. Because the agent Z can control agent X which controls agent Y, Z can indirectly control Y; this is referred to as the transitivity of control. This is true even when X is self-controlled: one can control things which are self-controlled by controlling "the states of the world that cause it, in controlling itself, to act" (*Elbow Room* 56). An example of this transitivity of control is how one is able to control a computer which controls another thing, and this same idea can be applied to the controlling of self-controlled humans. Thus, though self-controlled, due to the transitivity of control, an agent can be controlled by other controllers as well as oneself.

One way for B to escape being controlled is to make its activities random so that A cannot identify its states or limitations. This does not have to be mathematical randomness because pseudo-random series are often just as inscrutable. Dennett writes that "a certain amount of arbitrariness, of informational insensitivity… is an essential feature of time-pressured rationality" (*Elbow Room* 70). This is because, while self-evaluation tends to lead to improvements in the character of the agent, one is always at the mercy of the deliberation process, where deliberation is the act of making a decision based on the results of one's past decisions. Knowing this, one can work to become a better deliberator, which includes overcoming the debilitation that can be caused by an over-abundance of deliberation or self-evaluation. Sometimes a random choice is the best choice.

Ideally, one wants to be as immune from negative manipulation and as sensitive as possible to problems which may arise with its control so that it will have the 'elbow room' necessary for facing the world. In an example of Dennett's a pilot is told that a storm is up ahead that is so severe that if the pilot enters it he risks losing control of the plane. The pilot does what any rational person would do and steers his plane away from the storm. This shows that "...the pilot not only strives to control the plane at all times; he also engages in meta-level control planning and activity" by taking steps to avoid situations that will diminish his control (*Elbow Room* 64). For this ideal to be realized, one prefers to have more options than fewer when deliberating. This preference for greater spontaneity is partially dictated by reason: as mentioned earlier, a dose of randomness in one's decisions is healthy and sometimes optimal as it limits the likeliness of losing control to another agent and saves time which would have otherwise been wasted deliberating unnecessarily. This reveals the conflict between the ideal of spontaneity and the ideal of rational deliberation to be an illusion. Randomness, in other words, can be rational.

Determinism and control are not mutually exclusive. As Dennett writes, "the thesis of determinism carries no implications about how in particular this world will arrange itself" (*Elbow Room* 73). In other words, determinism does not imply that a device must act in a certain--say, restrained--way. Thus, a deterministic device can be a self-controller and can also evade another self-controller from learning about its states and controlling it, all while maintaining its determinism.

Part II: The Self as a Controlling-Agent

The self is either an acting-agent or a conduit of causation, but cannot be both. Dennett defines the self he is interested in as the former: "a locus of self-control" and "the sum total of the parts [one] control[s] directly" (*Elbow Room* 81-82). The growth of this self is closely related to the increase in one's capacity for self-control over time. When one is a baby, for example, one is barely a self. As time passes and the baby grows, it experiences more situations where it must analyze the success of its actions over time and make decisions based on what it learns: this is the process of deliberation. Thus, through deliberation and, as the reader will see, with some luck, one can achieve a high order of self-control, and, as Charles Taylor says, define oneself in the process of creating oneself. This developmental process based on self-control plateaus at different times for different people, based on their skill and luck; in this sense, self-formation is unlike any other material formation and must be thought of more like an art than a calculable science.

There is an intricate relationship between skill and luck: the more skillful one is, the less one's successes are considered to be due to luck. This is because skill corresponds directly to increased process control. There are two types of luck relevant when considering an agent's circumstances: initial luck and luck on a particular occasion.

One's initial luck is based on the faculties one is born with; this, at first, seems unfair, and so the issue of fairness must be considered. Kurt Vonnegut's short story "Harrison

Bergeron" examines a world where everyone is made equal through handicaps required by law for those with unfair advantages such as heightened intelligence, beauty, speed, or strength. This world without competition is certainly more fair for everyone, but it is also extremely sad and bleak--a fact the people in it are too dumbed-down by their handicaps to notice. Dennett believes that the apparent unfairness of some people starting their lives with initial advantages is irrelevant. He explains this through his example of runners in a race who are made to start out at slightly different distances so that some are a tad closer to the finish line than others. In a 100-meter dash this would make all the difference; life, however, according to Dennett, is more like a marathon: it is not a race to a finish but is instead the developmental process of a highly-functioning, self-controlling agent.

Dennett says it is important for one to not think of everything in terms of only luck but also leave 'elbow room' for skill. This is because, in a system where all that happens is based on luck and one is aware of this, it is impossible to act as if one is causing change. Such a luck-centric system cannot be true because humans are constantly deliberating to act in ways which will 'make a difference,' and then executing their decisions. If skill also exists then it is possible to say that "luck averages out and skill will tell in the end" (*Elbow Room* 97). Deliberation is an example of a skill gained over time that does not require luck. As will become clear through the explication of Dennett's thesis--because deliberation implies self-control, and because one is responsible for what one controls--one is responsible for one's actions when skill is involved. The 'elbow room' that exists for skill is thus required to define one as a responsible self-controlling agent.

Part III: Deliberation as the Method

Deliberation, as described earlier, is the act of analyzing the success of one's actions and using the results of past decisions to decide on future ones; as such, it implies self-control. If determinism were true, it would seem that the time one spends deliberating is a waste: the two appear incompatible, even absurd. Most people, that is to say, believe that if determinism were true, then no matter how effective determined deliberation might appear to be it would still not be *real* deliberation because the outcome had been determined from its onset. But Dennett says this is not so. He points out that people deliberate daily but that this does not automatically imply that determinism is false when he says, “what is actual is possible, but this manifest actuality of deliberation hardly shows that determinism is false” (*Elbow Room* 102).

If no deliberation was real or effective*,* the rational response would be to cease deliberation. But, as Peter van Inwagen remarks, one who did not deliberate “would either move around in random jerks or scuttles, or would withdraw into catatonia,” and clearly this is not what humans do, though Dennett jokes that perhaps this is what happened to the trees: “perhaps in the olden days trees scampered about, preoccupied with their projects, until the terrible day when they saw the light [that is, when they realized the futility of their frantic deliberation] and had to take root and ‘vegetate’!” (*Elbow Room* 104)

Dennett explains that one mistake that must be cleared up is the confusion between fatalism and determinism. Fatalism is the idea that there is a pattern in human affairs that will impose itself no matter what one tries to do. Local fatalism is an occurrence of fatalism, defined by Dennett in his paper “I Could Not Have Done Otherwise--So What?” as a “particular circumstance in the relevant portion of the past which ensured that the agent would not have done otherwise (during the stretch of local fatalism) *no matter what he tried, or wanted, to do*” (“I Could Not Have Done Otherwise--So What?” 554). Instances of local fatalism--that is, specific instances of fatalism--can occur in either deterministic or indeterministic worlds. Instances of local fatalism do exist but discovering that one is in “such an eddy of local fatalism” is not the same as discovering that the world is determined (*Elbow Room* 105). To explain this, consider the example of local fatalism Dennett provides in “I Could Not Have Done Otherwise--So What?” of an agent being locked in a room. In this case, no matter whether determinism is true or not--specifically, whether the agent was determined to be in the room or not--the agent cannot do otherwise but be in that room. Whether the agent’s actions in the room are determined or not is also irrelevant to the fact that the agent is stuck in the room. A perhaps more interesting and sophisticated manifestation of local fatalism can occur when an agent succumbs to its temptations and, due either to its own self-deception or weakness of will, cannot control its actions.

One must take extreme care not to confuse determinism with local fatalism, as the former does not imply the latter. While instances of local fatalism provide situations where deliberation is futile, such cases are, for the most part, abnormal. Nonetheless, the distinction between fatalism and determinism is integral to Dennett's "could have done otherwise" argument, which is addressed later on.

Still, though deliberating in a fatalistic world is ineffective, and though it has been made clear that fatalism is not the same as determinism, deliberating in a determined world seems just as fruitless. It leads one to ask: what would a deterministic deliberator even look like? And in response Dennett, extending an idea of Wilfrid Sellars's called the *manifest image*, has prepared one a portrait. However, to understand the necessity behind Dennett requirements of the perfect deterministic deliberator, call it C, it is first prudent to define one's manifest image. The deliberator-agent needs to have a method for collecting and sorting the information it receives about the world; this conception of the world held by the deliberator-agent is what Dennett calls its manifest image. One's manifest image is at a low-level made up of either things which are background conditions and thus cannot be changed or things which have multiple possible states and can be changed. At a higher level, the manifest image encapsulates things so that "among these possible states of things, some are simply unpredictable, some are reliably predictable... some are indirectly controllable by (the effects of) actions of the deliberator, and some are directly controllable by the deliberator" (*Elbow Room* 111). The point of this data management, and what this all ties back to, is the deliberator's direct ability to make good control decisions.

So on to the description of the perfect deterministic deliberator, C: it would be able to foresee events occurring in the very near future and its mind would be capable of sorting the "bad" or irrelevant past data from the "good" or valuable past data to use when deliberating about future data. C would be able to keep track of the changing features in its environment and be able to sort the non-chaotic features into trackable categories from the chaotic or unpredictable features which are still important. There are two varieties of events which are unpredictable to C: the first includes events related to the outcome of C's deliberations that are beyond its scope, and the second includes events concerning things on which C has not deliberated which are "unpredictable in practice" (*Elbow Room* 112). An example of the first type of unpredictable event is applying for a job and waiting for a response, while an example of the second is the occurrence of a sudden storm. Note that everything up until this point describes C's manifest image. The final requirement for C is that it be capable of self-prediction, though this has its limits; what the deliberator will decide to do arises from many possibilities, and it is often very difficult for it or anyone else to predict which ones will occur to C while it is deliberating. Here the reader must pause to note that every aspect of the deterministic deliberator C as set forth by Dennett is part of the method of human deliberation; the perfect *in*deterministic deliberator, D, would have the same character and properties, its only difference being that it is not determined.

Dennett writes, "the manifest image of any deliberator will include a partitioning of things into some that are to emerge as the results of the deliberator's deliberation--things that are thus 'up to' the deliberator--and things, predictable or not, fixed or not, that are not up to the deliberator." In more relevant terms, Dennett is saying that one's manifest image is made up of things which are in its control and things which are not in its control. As long as the agent accepts that some constraints are not under its control and recognizes the things that are in its control, it acquires the 'elbow room' it needs for deliberation. As noted, C's and D's modes of deliberation are identical. Dennett's conclusion about the manifest image reiterates this when he writes that it is valid for "any deliberator," and states that uncontrollable events are either "fixed or not," which means that events not under the agent's control can be either determined or undetermined. This follows easily from his ideas of control established earlier where it is shown that determinism and control are not mutually exclusive: a deterministic agent can still be a self-controller. (*Elbow Room* 113)

Although C's actions are determined, C cannot consistently predict the outcomes of its deliberations due to the limits of its self-prediction capacity. Thus, C does not know which decisions it is determined to make. Because C's determined character does not erode its ability to self-control, and because C recognizes the components of its manifest image, C can still deliberate. In fact the best chance C has of arriving at the outcomes it desires, whether these outcomes are determined or not, is through deliberation.

Dennett explains that to deliberate well, one must act as if the world had *real* opportunities. He defines an opportunity as "a 'chance' for an agent to 'do something' that will 'make a difference'" (*Elbow Room* 115). Only agents can deny other agent opportunities. This denial of opportunity can be done in two ways: either by creating islands of local fatalism for the other agent through brute force restraints--to use the example from earlier, this can be done by locking the agent in a room-- or by keeping the other agent in the dark about opportunities it may have. This latter instance of denied opportunity is called a bare opportunity because it is unimagined and unrecognized as an opportunity to the agent who is unaware of it. Bare opportunities are plentiful, but one wants more than this; one wants to be aware of opportunities in time to act on them. And so a real opportunity is defined as an opportunity that the self-controller is made aware of in time to act. While at first it might seem contradictory to be able to "made a difference" in a determined world, Dennett asserts that determinism does not rule out opportunity.

As usual, it is easiest to understand Dennett through metaphors, so recall C, the perfect deterministic deliberator, and its friend[3] D, the perfect indeterministic deliberator, and assume them to be imperfect so they both fail to act on their own separate real opportunities. Possible reasons for their failure include deliberating the wrong way, taking too long to come up with a decision, or not prioritizing a piece of data correctly when deliberating; the specific reason is unimportant in this case study. Most people would be inclined to say that C, being determined, never really had a chance--that is to say, C never had a *real* opportunity. D, on the other hand, is undetermined and so it seems that things are more fair for it-- it seems that D's opportunity was *real*. Dennett claims that this supposed difference in fairness for C and D is an illusion.

To explain the fallacy of there being unfairness for C and fairness for D, Dennett offers another metaphor, this time of two lotteries the same in every respect except: in Lottery 1 the tickets are sold and the winner is chosen and announced after all the ticket stubs have been suitably mixed; in Lottery 2 all of the ticket stubs are mixed first, the winner is chosen and kept secret, then the tickets are sold and the winner announced. One's chances of winning in Lottery 1 are the same as in Lottery 2; the timing of the selection of the winner does not make a difference

---

[3] Can they really be friends? See forthcoming "On Being Friends with a Determinist" by Cadareanu.

to one's chances of winning. Fairness does not mean that everybody wins: it means everyone has a chance to win. Therefore, in this example both lotteries are equally fair. Dennett says that ordinary people know that is true, as evidenced by the fact that lotteries of both kinds exist and are popular, but philosophers seem to have convinced themselves that "without a continual supply of genuinely random *cruces* to break up the fabric of causation, there cannot be any real opportunities or chances." ("I Could Not Have Done Otherwise--So What?" 564)

Dennett likens Lottery 1 to an undetermined world and Lottery 2 to a determined world, and says that one's chances are the same in either one. Of course, some people will have more initial luck, but this is true whether the world is determined or not, as has been explained antecedently. In this same sense, some people have more opportunities than others, and not all opportunities are the same; the opportunities that matter for a person are based on the features of one's particular manifest image.

Roderick Chisholm claims that acts with "'sufficient causal conditions... are not avoidable.'" According to Dennett this statement implies that if determinism is true then, "since all our acts will have sufficient causal conditions, no act of ours is unavoidable." (*Elbow Room* 123) Chisholm's claim, however dangerous it appears, can be disregarded on the basis that only instances of local fatalism are truly "unavoidable," or inevitable. One cannot, as the idiom goes, "change the past," and in this same sense one cannot "change the future" either: the past is what has already happened and the future is what will happen next. Whether or not the future is determined is unknowable to humans. But this does not make one's future *inevitable*: the future--even if it is determined--always appears open because one cannot see what it will be. So one can "change the future" from what one thought it would be into something else. And this is the definition of free will: free will is one's capacity to see one's probable futures in time to deliberate on them and take steps, through each decision, towards the future one desires.

Part IV: "Could Have Done Otherwise"

One of Dennett's strangest but surprisingly strong arguments is against what he calls the "could have done otherwise" principle. This principle is often considered one of the necessary conditions for holding an agent responsible for an act one has committed and is as follows: if an agent could *not* have done otherwise, it cannot be responsible for its actions.

An agent, call it E, makes a decision to do something considered bad--for example, murder another agent--and is subsequently put on trial. The first question most people ask when judging E's moral responsibility in the case is "could [it] have done otherwise?"[4] At first glance one might think this is because CDO appears to be questioning the agent's determinism or indeterminism in that instance; this might be the goal of the CDO principle due to the usual association of indeterminism with responsibility. Dennett however does not care about the intent of CDO; he writes that "*whatever* 'could have done otherwise' actually means, it is not what we are interested in when we care about whether some act was freely and responsibly performed" (*Elbow Room* 131-132).

Harry Frankfurt is one of the few philosophers who also challenges the CDO principle. He does so with the following intuition pump:

> Jones hates Smith and decides, in full possession of his faculties, to murder him. Meanwhile Black, the nefarious neurosurgeon… who also wants Smith dead, has implanted something in Jones's brain so that *just in case Jones changes his mind* (and chickens out), Black, by pushing his special button, can put Jones back on his murderous track. In the event Black does not have to intervene; Jones does the deed all on his own. (*Elbow Room* 132)

In this case, Frankfurt thinks Jones, though he really could *not* have done otherwise, deliberated and made the decision to kill Smith without (during the actual confrontation) any intervention by Black and so is completely responsible for his action. Dennett agrees with Frankfurt's conclusion that CDO is not relevant in this case, but thinks such an example leaves room for argument and wants to show that CDO is not relevant *in any case.*

Dennett points out that one does not withhold judgement based on whether the agent on trial CDO or not: a person who one thinks could do no other is still blamed and praised for his or

---

[4] "Could have done otherwise" is colloquially referred to as CDO in this paper.

her actions. Following this train of thought, one does not need to know if a person's actions were determined or undetermined to judge him or her, and so CDO does not seem like a useful strategy for determining moral responsibility. Another example of CDO's inadequacy is its inconsistency in use: one does not ask CDO when another does a good deed. An example of this is as follows: agent G gets agent F agent F's favorite ice cream. It would be extremely odd, even bordering on inappropriate, for agent F to bring up the CDO principle in such a case to determine whether or not agent G was truly responsible for getting agent F the ice cream, and thus determine if agent G deserves being thanked.

Upon further inspection of the CDO principle, one notices that CDO's inquiry into causation ends with the decision of whether local fatalism was present or not. From this, one finally derives the assumption of CDO which renders it mostly useless in judging moral responsibility: CDO assumes that local fatalism implies determinism, and the lack of local fatalism implies indeterminism. As explained earlier, these are fallacies: local fatalism is "entirely neutral between determinism and indeterminism" ("I Could Not Have Done Otherwise--So What?" 555). In terms of CDO, this means that if local fatalism is true then the agent really could *not* have done otherwise, independent of determinism being true or not.

The CDO principle is almost always applied in the exact same case; that is, if the situation were to be repeated exactly the way it occurred in the first case, then the CDO principle is addressed. A resolution cannot be found to the CDO principle in an exact case because a human agent is constantly evolving through its deliberations and will never be in the exact mental state as during the first case, if somehow all the external details of that first instance could even be replicated. In similar and not-exact cases--except in the case of local fatalism being true and causation ending there--the CDO principle can never be resolved seriously because a human agent is too intricate for one to be able to pinpoint that agent's causes for doing a certain thing. And even supposing somehow a resolution of the CDO principle in an exact and particular case could be found, it would not make any difference to what happened.

Besides the fact that CDO principle can almost never be resolved seriously, the question of CDO in the exact same case provides an uncomfortable conclusion about responsibility: if indeterminism is true at the subatomic level, it is possible that all of one's macroscopic actions are the effects of this indeterminacy, and it is just as possible that one's macroscopic decisions are determined. It is extremely unlikely that one could ever find good evidence of indeterminism at the subatomic level causing an event at the macroscopic level, so one can never know for sure if an act is determined or not. Based on this conclusion, if the CDO principle were applied in the exact same case this would imply that responsibility hinged on whether a decision was already determined. This means that one could never know if anyone has ever been responsible because one can never know for sure if an act is determined or not.

Dennett says that if the murdering agent were to find itself in a case similar to the previous one where it did action A (murder another agent), and in this new case it could do no other than action A again, then all this shows is that action A was the best choice in the case. If the agent did the same in all similar cases, this only strengthens the idea of action A being the best choice. Believers of the CDO principle, however, want to argue that "a genuinely free agent… must be more volatile" and claim that the agent's decision if chosen repeatedly in many similar cases shows that the agent is zombie-like and not responsible for its action. Dennett says if the agent were to choose the best option each time, all that this means is that it is rational: "the general capacity to respond flexibly in such cases does not at all require that one could have done otherwise in the particular case or in any particular case, but only that under some variations in the circumstances--the variations that matter--one would do otherwise." ("I Could Not Have Done Otherwise--So What?" 556)

A CDO believer might still think that if the agent was determined to murder, then it "*had no chance not to do it,*" but Dennett says such people do not understand that whether or not one was determined does not affect one's chances ("I Could Not Have Done Otherwise--So What?" 564). Recall the previous conclusion about chance and determinism: a deterministic and indeterministic world are equally fair, so a determined agent in the former has the same chances as an undetermined agent in the latter; their opportunities are different but this is because they are based on their different manifest images.

Dennett maintains that as long as there was no blatant local fatalism in play, whether or not it was determined to act, the agent that did the bad thing and was put on trial is responsible for its decision. Figuring out if the act was determined or not is physically impossible, and even if one could somehow know this, the thing still happened because of an action deliberated and decided on by the agent. Anyway, because people "are designed to be so sensitive to the passing show that they never can be in the same microstate twice," applying the CDO principle does not result in a deeper knowledge of the agent's character because the micro-causation of a human being in some particular circumstance will never occur again ("I Could Not Have Done Otherwise--So What?" 559). In other words, deliberating the causation that resulted in an agent acting in a certain way in a particular instance does not determine that agent's character, as one's character is based on more than a single action.

Dennett claims that what one is really looking for when one appears interested in the CDO principle is character assessment. One wants to know if the murdering agent's act of murder was a fluke or a telling trend. Whether or not the agent is determined is not important from that point of view, but it is important to analyze the agent and for the agent to analyze itself so as to interpret its actions, learn about itself, and "do better" in similar circumstances.

Part V: The Intentional Strategy for Deliberating

Dennett's intentional stance or strategy, which he explains in the paper "True Believers: The Intentional Strategy and Why It Works" as well as the first chapter of his book *Brainstorms*, is one of his most interesting ideas. It is a strategy for predicting the behavior of an object or system to be used if the physical[5] strategy is not relevant, and the design[6] strategy is not viable (as when the object or system being analyzed does not have designed behavior). The intentional strategy works as follows:

1. Treat the object as a rational object (or system).
2. Determine what beliefs the object ought to have.
3. Determine what desires it ought to have.
4. Predict that the object will act to satisfy the desires it has based on its beliefs.

---

[5] The physical strategy is as follows: 1. determine the object's physical constitution; 2. determine the physical impingements on the object; 3. use the laws of physics to predict the object's output.

[6] The design strategy is as follows: 1. ignore the details of the object's physical constitution; 2. assume the object has a design; 3. predict that the object will behave as it was designed to behave.

An object or system whose behavior is well-predicted by the intentional strategy is called a "true believer," and its status as such is, according to Dennett, objectively so. Rationality is the assumption that the object which has beliefs and believes the implications of those beliefs does not believe any contradictory beliefs. Usually, all that is needed for an object to develop true beliefs about a thing is being exposed to the thing over a reasonable amount of time. One can assume that the object ought to have desires that are good for it; for a person, as an example, some basic desires include food, shelter, and comfort.

The intentional strategy works successfully on a variety of systems, both living and not (as in the case of a computer), and it provides one with a predictive power otherwise unavailable. Clearly the strategy cannot be used to predict very specific things, as it is practically impossible to know the specifics of another system's beliefs or desires, but Dennett paints this as a strength when he writes, "it is this neutrality with regard to details of implementation that permits one to exploit the intentional strategy in complex cases." ("True Believers: The Intentional Strategy and Why It Works" 7)

It is possible to use the intentional strategy on lower-level systems; an example of this is using the intentional strategy to predict the behavior of, say, a folding chair. This would be as follows:

1. Treat the chair as rational.
2. What beliefs does a chair have? Probably none.
3. What desires does a chair have? Maybe to remain folded.
4. Prediction? It will remain folded.

This was successful in predicting the folding chair's behavior, but it did not provide any insight into the object that was not already known, and so it was a waste of deliberation.

Robert Nozick's objection to Dennett's intentional strategy and Dennett's nimble response serve as a backdrop for the consideration of its relevance in the discussion of deliberation. This objection is (of course) given in the form of a metaphor: Nozick asks the reader to suppose that Martians significantly smarter than humans would not need to take the intentional stance to predict human behavior in all its detail and thus, from their point of view, humans are not true believers. This means that the human status as true believers, if Nozick is right, is subjective. In response, Dennett insists that while fatalists are wrong to think that the apparent patterns in human affairs will force themselves upon the world and that humans "can do no other," they are right about the patterns; he uses this idea of patterns in human affairs to argue against Nozick. After acknowledging Nozick's criticism as admissible, Dennett explains that the Martians "would be *missing something* perfectly objective: the *patterns* in human behavior that are describable from the intentional stance, and only from the stance, and which support generalizations and predictions" ("True Believers: The Intentional Strategy and Why It Works" 7). Thus, Dennett believes it is impossible to avoid the intentional strategy with respect to self-controlling agents such as humans.

These patterns in human behavior are what one is interested in when one considers the CDO principle: these patterns are what emerge from character assessment. An agent is characterized in terms of its beliefs, desires, and intentions; these are all the parts required in the intentional stance, so it would seem that through the use of the intentional strategy one could successfully deliberate about oneself or other selves like it. The intentional strategy, then, is relevant to the discussion of deliberation because it *is* one's main method of deliberating about oneself and other agents.

Part VI: Where All of This Gets Us: Responsibility

Dennett says that the idea of responsibility as present in everyday society is that "I *take* responsibility for anything I make and then unleash upon the public" (*Elbow Room* 85). This means that one is responsible for what is within one's control. But, according to Charles Taylor, claiming authorship for one's choices is not enough for moral responsibility. This is because responsibility must be *taken*, whereas authorship is hardly ever questionable; in other words, an agent would have a hard time proving it did not create something it created, but could very easily not take responsibility for the thing created. Taylor writes, the "ground of responsibility… comes from the particular nature of the way in which we accomplish radical re-evaluation," and it is through the success of this re-evaluation process that the agent defines itself and its values; Dennett agrees wholeheartedly (*Elbow Room* 90).

Consider the point of *holding* someone responsible: why is it compelling to judge and punish those who commit crimes? Maybe, as Friedrich Nietzsche might argue, humans are naturally violent beings. Dennett supposes there are more reasonable responses: for example, crimes are harmful to society and by threatening punishment it is possible to decrease their frequency. This is fine, but it does not explain why one would want to take responsibility for oneself--for example, what benefit could result from holding oneself responsible in an instance where committing a crime would be beneficial to one, if not to society?

To understand the necessity and one's desire for moral responsibility, consider Paul Gomberg's idea that "by holding someone responsible and acting accordingly, we may cause him to shed an undesirable trait, and this is useful regardless of whether the trait is of his making" (*Elbow Room* 163-164). Now recall Dennett's claim that what one is really interested in when he asks CDO--which is used to determine one's moral responsibility in a certain case--is understanding the patterns in human behavior present in one's manifest image. One is specifically interested in the patterns which emerge from the results of one's deliberation.

Susan Wolf writes that "an agent can be both determined and responsible insofar as he performs actions that he ought to perform: if an agent performs a morally bad action, on the other hand, then his action cannot be determined in the appropriate way." In this claim, Wolf seems to suggest that for a morally bad action to occur, there must be something criticizable with the reasoning that led to the morally bad action. While Dennett agrees that responsibility does not require an indeterministic system, he disagrees with Wolf's idea that morally bad actions are the result of bad reasoning because reasoning cannot be perfect. Dennett writes, "the *best possible* designs, given the constraints of finitude and time pressure, would have to include some measure of arbitrariness and wise risk taking." (*Elbow Room* 164)

So because deliberation includes some risk-taking and arbitrariness, it is a process that can (and often does) result in mistakes, including morally bad actions. One wants to limit these mistakes, and one can do this through the same paths as sketched in Dennett's account of the rationale of punishment by law: "by somewhat arbitrarily holding people responsible for their actions and making sure they realize that they will be held responsible, we constrain the risk-taking in the design (and redesign) of their characters within tolerance bounds" (*Elbow*

Room 165). The reason Dennett proposes an “arbitrary” scheme of responsibility is because in such a world everyone is held responsible at all times, so no one can object to the fairness of such a system of responsibility.

The problem with Dennett’s arbitrary idea of responsibility is that in an “arbitrary” system one is not *obliged* in any way to do anything (in this case, one is not obliged to hold oneself responsible). Dennett acknowledges this but maintains that without holding oneself responsible it is impossible for one to improve oneself.

Dennett writes: “in Sartre’s view, it is as if one could create oneself *ex nihilo*: as if, to borrow another familiar bit of philosopher’s Latin, the self at birth were a *tabula rasa* or blank slate” (*Elbow Room* 83). This is an odd reading of Sartre and a different one could do Dennett’s idea of responsibility more justice.

Sartre outlines his early ideas of absolute responsibility in his papers “Freedom and Responsibility” and “Existentialism is a Humanism,” the main points of which follow. First, freedom as understood and espoused by Sartre is a power which cannot be affected by people and through which all one’s willing occurs. Second, he maintains that each human being is responsible for all of humanity’s actions, so much so that the human being must claim them as his or her own; Sartre writes, “everything which happens to me is mine... I am always equal to what happens to me *qua* man.” (“Freedom and Responsibility” 53) This seems an extremely constraining and hefty vision of responsibility, but he does not stop there. Next, he claims that, though the human being is at first not responsible for its facticity[7], its reaction to its birth as implied by continuing to live assumes full responsibility of its facticity and claims the facticity as its own.

Sartre provides an example of one’s absolute responsibility when he writes that a coward is completely responsible for its cowardice because it made himself a coward through its actions. He then says that most people would prefer to be born heroes, but in existentialist thinking “there is always a possibility for the coward to give up cowardice and for the hero to stop being a hero” (“Existentialism Is a Humanism” 227). From this, one can see that, however constraining

---

[7] Facticity is one's presence in the world as limited by the facts of his or her existence; this is similar to Heidegger's “throwness.” That is, facticity is one’s abandonment into being, which is its birth.

Sartre's absolute responsibility might seem, absolute freedom is directly implied through it, and this absolute freedom is realized through every one of the human being's actions.

The human being cannot explain its existence, or the existence of others: "facticity is everywhere but inapprehensible; I never encounter anything except responsibility," but the human being is still compelled to understand the meaning of its being, and this leads it into despair ("Freedom and Responsibility" 58). The only solution to this is for the human being to, in anguish, fully comprehend that it *is* being and is "thrown into a responsibility which extends to [its] very abandonment." Once the human being claims this responsibility, as Sartre writes, "[it] is no longer anything but a freedom which perfectly reveals itself and whose being resides in this very revelation." (Sartre 59) Sartre would never even consider the CDO principle: for Sartre responsibility is an eternal condition of being human and there is no instance where one could ever not be held responsible.

According to Sartre, it is an existentialist's natural duty to judge those who hide from total freedom and those who deny it and try to show that their existence is necessary. One can pronounce a moral judgement in such a case because, while freedom as a definition of the human being does not depend on other persons, the human being is a free being who cannot do anything but will its freedom. This means that, when put in a social situation, the human being, who cannot do anything but will freedom, cannot *not* will the freedom of others. Besides this, however, an existentialist cannot judge what the "right" thing is for another person in a certain situation. This is because the human being, if it acts through freedom, is unable to choose anything but what is the best for itself. Note that what is best for all other beings is intimately connected with what is best for a specific human being, and this can lead the human being into anguish and despair because he or she must choose for everyone.

Dennett calls Sartre's idea that one chooses oneself "ultimately incomprehensible." He says that Sartre requires such an exaggerated idea of freedom because he is trying to combat the argument that "unless one were absolutely responsible for oneself, one could not be responsible at all." This is an interesting criticism considering Dennett's system of responsibility where everyone is to be held arbitrarily responsible for all of their actions. The main problem Dennett has with Sartre's idea of absolute responsibility is Sartre's declaration that one can claim responsibility for one's very facticity. (*Elbow Room* 83)

In "Freedom and Responsibility" Sartre claims that one is responsible for everything except one's own responsibility, because one cannot be the foundation of its being. He explains that the fact of one's birth "never appears as a brute fact but always [appears] across a projective reconstruction of [its] for-itself" ("Freedom and Responsibility" 57). This means that one's birth is always associated with a reaction and is not just an objective fact or idea: for Sartre, one's life begins with a choice to react to one's facticity, which is one's birth.

It is easy to see how Dennett reads Sartre: for existentialists "existence precedes essence" and this seems to imply that one defines oneself completely on the basis of one's absolute freedom. Thus, while it is viable to think of Sartre's idea of the reaction to one's facticity as a "first choice"[8] written on the *tabula rasa* of one's life, this does not get Dennett any further in defining his own system.

In Dennett's system everyone is held responsible for their actions because responsibility leads to better deliberation. Sartre's responsibility is absolute and so applies to everyone as well, but his idea of responsibility carries more weight than Dennett's. But how can this be, when both ideas of responsibility hold everyone responsible? Recall that deliberation is not perfect, but is instead a gradual process tending towards perfection. Dennett's arbitrary responsibility is the same; one goes from being a being with no responsibility to being a being with responsibility--and besides that, a responsible character--through the gradual acquisition of responsibility. This means that for Dennett the same things are not expected of a baby as of a fully-formed, well-functioning adult. It is not fair to say that Sartre would expect the same things from a baby as an adult, but for Sartre absolute freedom and the implied absolute responsibility are true for all humans since their facticity. Does this mean that Dennett considers babies to have 'less free will' than adults? In a sense, yes. For Dennett, having 'free will' is not the same as not

---

[8] On the other hand, perhaps one's absolute freedom makes it so one's *tabula rasa* is not existentially void--that is, one can argue that for Sartre because birth is always associated with a reaction and is not just an objective fact, one does *not* start out with a *tabula rasa* because the choice to react to one's facticity is already written on one's slate.

being determined; the former is a matter of control--of having as much ‘elbow room’ as possible--and a baby certainly has smaller elbows and less control than an adult.

In “Existentialism is a Humanism” Sartre writes, “if indeed existence precedes essence [that is, if existentialism is correct], one will never be able to explain one’s action by reference to a given and specific human nature; in other words, there is no determinism--man is free, man *is* freedom” (“Existentialism Is a Humanism” 223). From this it is obvious that Sartre does not share Dennett’s views that determinism and control are compatible.

Dennett’s conclusion is that one wants as much ‘elbow room’ as possible to deliberate between multiple resolutions and make a meaningful choice. There is no way of guaranteeing that one will have any ‘elbow room’--that is, no way of guaranteeing an indeterministic world--but believing in free will is one of the necessary states of having it. This means there is no reason for one to *not* think and act as if one had free will. The implications of Dennett’s compatibilist view require that some traditional beliefs about blame and responsibility be renounced. Dennett’s main example of this is the basis of culpability on total, before-the-eyes-of-God guilt; he claims that the integrity of the institution of moral responsibility is better off without the traditional idea of God.

In the sense that there are free wills worth wanting, for Dennett there are also freedoms not worth one’s time; these have been mentioned in this paper but were not highlighted as such until now. Two examples of unwanted freedom include the absolute freedom proposed by Sartre and the freedom to choose arbitrarily at all times. The latter was mentioned in the discussion about the CDO principle where Dennett says it is unnecessary to choose differently in a similar case than in the first case for one’s choice to be considered an action of free will; random deliberation is sometimes necessary--such as in time-sensitive situations--but if all deliberation were arbitrary one would not be rational and one’s moral responsibility would never develop because there would be no point to it.

Criticism

Dennett claims, at the start of his book *Elbow Room,* that the only reason the free will problem appears such an ardent matter to one who would otherwise not be interested in philosophy is because philosophers and writers have created so many uncomfortable or even downright terrifying metaphors for an existence without free will. He calls these metaphors ‘bugbears’ and claims that they must be taken head-on if one hopes to ever dispel the fears they instill. An example of a bugbear is what Dennett calls ‘the Bogeyman’: an evil agent who competes with one for control of one’s body to use for its own malicious purposes; another example is of the ‘Invisible Jailer’ who is imprisoning one in a not-so-obvious jail. This latter bugbear has been explored in many science fiction media, including the 1998 movie *The Truman Show* where the protagonist Truman Burbank’s life since his birth and adoption by a T.V. production company has been staged and recorded for the viewing pleasures of strangers. At first, Truman is unaware that he is the star of a T.V. show and that his life is fake, and in this sense he is imprisoned in an ‘invisible jail.’

Dennett cannot prove that none of the bugbears he lists exist, but he says there is no evidence to suggest any of them do, either, and “a closet with a ghost is a terrible thing, but a closet that is just like a closet with a ghost (except lacking the ghost) is nothing to fear.” To counter these bogeymen, one must first recognize them and then “check to see if [these] scary agent[s] are really doing all the work,” which one can do if one is able to identify the bugbear’s methods for instilling fear. (*Elbow Room* 10)

A bugbear works as follows:

1. The bugbear starts off providing a simple case of something awful (such as being controlled by another agent).
2. Next, a potential similarity to one’s own case as a self-controlling human agent is brought up, though lacking a significant amount of complexity.
3. “Thanks to the similarity chain,” this is awful.

This “deliberate over-simplification of tasks to be performed by the philosophers’ imagination” is a case of abusing intuition pumps[9] (*Elbow Room* 12). In the case of a bugbear, it is easy for one to derive an emotional intuitive judgement due to the simplicity of the case. One example of this is caused by the associated similarity of the fear of sphexishness described by Douglas Hofstadter and the invisible jailer bugbear. Sphexishness is a property based on the digger wasp

---

[9] Term coined by Dennett in *Consciousness Explained*.

Sphex ichneumoneus who at first appears organized and thoughtful but then is revealed to have no true mindfulness and be “driven inexorably into her states and activities by features of the environment outside her control” (*Elbow Room* 11). One wants one’s mindfulness to be real, and so one fears being sphexish. Thus, when one is given the example of the invisible jailer, the association between something feared--that is, the loss of mindfulness--causes one to derive a heartfelt intuitive judgement from the simplicity of the case--that is, one fears the bugbear and the associated example--instead of considering the actual content of the example where the bugbear is being used.

Intuition pumps are thought experiments which do not follow the scientific method that proves conclusions from given premises: “rather, their point is to entertain a family of imaginative reflections in the reader that ultimately yields not a formal conclusion but a dictate of ‘intuition’.” Sometimes their abuse is unintended, and intuition pumps are extremely useful tools for explaining philosophy on a basic level (to quote Dennett, they help people “see the forest and not just the trees”), but they are extremely dangerous when misused and not seen for what they are. One example of this abuse occurs when one derives an intuitive judgement from the simplicity of the imagined case instead of from the actual content of the example. (*Elbow Room* 12)

At the end of his book Dennett claims to have ‘exorcised’ each bugbear by showing that their existences were enabled by misused intuition pumps. The specifics of these exorcisms are not of interest here; instead, consider Dennett’s ideas as they have been exhibited in this paper. How many of them required intuition pumps in order to be credible? It appears as if the majority of his arguments are practically impossible to imagine *without* intuition pumps. Of course Dennett claims that his “own intuition pumps are designed to help” imagine the otherwise unimaginable (*Elbow Room* 170). This is true; for example, the conclusion that C, the deterministic deliberator, and D, the indeterministic deliberator, both had real opportunities and thus are treated equally fairly is very easy to imagine after considering the Lottery intuition pump provided. Before the Lottery intuition pump, however, one’s intuition was that C had an unfair chance compared to D; this shows how fluctuant and shaky one’s intuition is. Basing one’s philosophy as heavily as Dennett has on the use of intuition pumps seems to result in a system of thought with *more wobble* than appears at first, second, or even third glance. That being said, Dennett's compatibilist philosophy of free will seems to be deductively valid, though as of yet it is indeterminable whether it is sound.

Acknowledgement

The author wishes to thank her wise advisor, Dr. David B. Suits, for re-asking her questions back at her, and providing continuous support and criticism during this paper's various stages of development.

Works Cited

Dennett, D. C. *Elbow Room : The Varieties of Free Will Worth Wanting*. Cambridge, Mass.: MIT Press, 1984. Print.

---. "I Could Not Have Done Otherwise--So What?" *Journal of Philosophy*.81 (1984): 553-65. Print.

---. "True Believers: The Intentional Strategy and Why It Works." *Scientific Explanations: Papers Based on Herbert Spencer Lectures in the University of Oxford*. Ed. Heath, A. F. Oxford: Oxford University Press, 1975. 53-75. Print.

Sartre, J.P. "Existentialism Is a Humanism." 1946. Web. 15 October 2015.

---. "Freedom and Responsibility." Trans. Barnes, Hazel E. *Existentialism and Human Emotions*. Being and Nothingness. New York: Carol Publishing Group, 1990. 52-60. Print.

Bibliography

Chisholm, R. "Responsibility and Avoidability." *Determinism and Freedom in the Age of Modern Science*. Ed. Hook, S. New York: New York University Press, 1961. Print.

Dennett, D. C. *Brainstorms : Philosophical Essays on Mind and Psychology*. Bradford Book. MIT Press ed. Cambridge: MIT Press, 1981. Print.

---. *Consciousness Explained*. 1st ed. Boston: Little, Brown and Co., 1991. Print.

---. *Darwin's Dangerous Idea : Evolution and the Meanings of Life*. New York: Simon & Schuster, 1995. Print.

---. *Freedom Evolves*. New York: Penguin, 2004. Print.

---. *Intuition Pumps and Other Tools for Thinking*. First edition. ed. New York: W. W. Norton & Company, 2013. Print.

---. "Postscript: "Reflections: Real Patterns, Deeper Facts, and Empty Questions"." *The Intentional Stance*. Cambridge, MA: MIT Press/ Bradford Books, 1987. 37-42. Print.

Frankfurt, H. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy*.65 (1969): 829-33. Print.

Gomberg, P. "Free Will as Ultimate Responsibility." *American Philosophical Quarterly*.15 (1978): 208. Print.

Hofstadter, D. R. "Can Creativity Be Mechanized?" *Scientific American*.247 (1982): 18-34. Print.

Hume, David, Thomas Hill Green, and Thomas Hodge Grose. *A Treatise of Human Nature : Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects : And Dialogues Concerning Natural Religion*. 2 vols. London ; New York etc.: Longmans, Green and co., 1898. Print.

Sellars, W. *Science, Perception and Reality*. London: Routledge & Kegan Paul, 1963. Print.

Taylor, C. "Responsibility for Self." *The Identities of Persons*. Ed. Rorty, A.: University of California Press, 1976. 281-99. Print.

The Truman Show. Dir. Peter Weir. Perf. Jim Carrey, Laura Linney, Noah Emmerich, Ed Harris, and Natascha McElhone. United States: Paramount Pictures, 1998. DVD.

van Inwagen, P. *An Essay on Free Will*. Oxford: Clarendon Press, 1983. Print.

Vonnegut, Kurt, and Gregory D. Sumner. *Welcome to the Monkey House : A Collection of Short Works*. The special edition. Dial Press trade paperback edition. ed. New York: Dial Press Trade Paperbacks, 2014. Print.

Wolf, S. "Asymmetrical Freedom." *Journal of Philosophy* LXVII (1980): 151-65. Print.