A Buddhist Reductionist Theory of Moral Responsibility and Freedom

Sangyop Lee
Department of Philosophy, Seoul National University

"There are no moral phenomena at all,
only moral interpretations of phenomena."
— F. Nietzsche, *Beyond Good and Evil*

In this paper, I attempt to suggest a possible Buddhist answer to the problem of determinism and free will. The argument will be that a Buddhist who is committed to mereological reductionism can give a compatibilist account of the problem by drawing upon the distinction between conventional and ultimate truth and Frankfurtian approach to moral responsibility and free will.

I first introduce van Inwagen's renowned Consequence Argument for clarifying the problems at issue and briefly discuss possible options a Buddhist can take regarding the argument.

I then summarize some previous attempts to provide a Buddhist answer to the problem of determinism and free will, and lay out what I consider to be right or wrong about their arguments. The discussion mainly deals with Goodman's hard determinist approach, and Siderits' paleocompatibilism.

Then I will try to propose a possible understanding of free will that is compatible with both determinism and Buddhist mereological reductionism. Main theses I am going to advocate or argue are the followings:

(1) At the ultimate level, the problem of free will or moral responsibility does not arise.
(2) The Principle of Alternate Possibilities is false.
(3) What we in conventional discourse designate with the phrase "a morally responsible action of John" or "an action done of John's free will" is a whole made of parts, and its existence is totally compatible with determinism.

The Consequence Argument and Options for Buddhists

In his famous essay "An Argument for Incompatibilism" (1983), van Inwagen introduces a seemingly unassailable argument to demonstrate that if determinism is true, then people do not have free will, which is later known as the Consequence Argument. I am here going to roughly reconstruct his argument because I think doing so will help discern and introduce various issues I am to discuss in this paper.

First, let P0 be a conjunction of every facts about the world before J was born, and let P denote a conjunction of facts about the world at the time *t* when J did A, and let L be conjunction of laws of nature which would be equivalent to the law of dependent origination for Buddhists, then,

(1) If determinism is true, then the conjunction of P0 and L entails P.
(2) J's not doing A at *t* contradicts with P.
(3) So if J could have done not-A at *t*, then he either could have rendered P0 false or he could have made L false. (1, 2)
(4) By the definition of L, J could not have made L false.
(5) By the definition of P0, J could not have rendered P0 false.
(6) J is morally responsible for doing A (that is, J did A by her free will) only if J could have done otherwise than doing A

at *t*.

(7) If determinism is true, J could not have done not-A at *t*. (1, 3, 4, 5)

(8) If determinism is true, J does not have free will. (6, 7)

Although this argument seems valid and sound, there can be objections.

(1) is trivial because this seems to be the right definition of determinism. If someone would argue that P0 and L do not necessarily entail P, then it would not be a different interpretation of determinism but just an indeterminist theory. And any Buddhist who accepts the law of dependent origination to be ultimately true, such as Abhidharmikas or Yogācārins, would not disagree with this definition.

I think accepting (1) entails rejection of any agent-causation theory, because any agent-causation theory would imply that the conjunction of P0 and L does not necessarily entail P as it posits a moral agent to be the ultimate initiator of a causal series. Also since it is evident that any Buddhist would not consider this option seriously, it is not in this paper's interest to investigate a possible agent-causation theory.

(2) also must be accepted unless one intends to deny the principle of non-contradiction. I simply cannot do A and not-A at the same time, *t*.

Some Buddhists can argue (3) is problematic because it is violating the insulation between conventional discourse and ultimate discourse, (1) being essentially a statement in the ultimate level since it should be about causal series of ultimate *dharma*s and (2) being a conventional statement since it is speaking of a conventional concept "person". I take this to be the view of Siderits's Buddhist paleocompatibilism which I basically agree with.

The definition used in (4), that if a law could be broken, then it

would not be a law of nature, is agreeable with the Buddhist understanding of the law of dependent origination as one of the Four Noble Truths, which are universal and ultimately objective truths.

(5) might be problematic, because Buddhists think that there was an exclusive causal series before a person's birth that caused another causal series we now denote as that person in this life; in short, Buddhists believe in rebirth. But this would be appropriating a loophole in the argument rather than addressing the issue seriously, so I will not go further with this objection.

A Buddhist can also reject (6) and argue for compatibilism, that is, a Buddhist can reject the Principle of Alternate Possibilities and argue that even if persons can only perform what was causally determined to be performed, they are nevertheless capable of acting on their free will and thus are morally responsible. This basically is the option I am about to explore in this paper. Repetti also uses Frankfurt's refutation of PAP and his definition of free will as accordance between volition and meta-volition to suggest a Buddhist compatibilist account to the problem. But I think similarity between my and Repetti's interpretations stops here.

Another option a Buddhist can take is to accept the whole Consequence Argument to be valid and thereby argue that Buddhists, as long as they consider the law of dependent origination to be ultimately true, should also embrace the fact that there is no free will in whatsoever sense, which is Goodman's hard determinist interpretation.

Buddhist Hard Determinism

Goodman, in his "Resentment and Reality: Buddhism on Moral

Responsibility" (2002), accepts the Consequence Argument, and claims that, "Anyone who does believe in free will has to accept a complete mystery which no one can think of any way to justify, or even to understand."

He goes on to point out our reluctance to accept a more rational option, the hard determinist view, stems from our intuitive concern that it would render everyday social practices impossible. We reproach morally bad behavior and praise morally good behavior, and we also often seek justice by institutionally punishing people for their immoral actions. But if we accept hard determinism, these social norms must be discarded because we can not hold a person to be morally responsible for her behaviors, for she was not the one who ultimately caused that action, and because she could not have done otherwise, because she did not really have a choice but to do it.

However, Goodman raises an interesting question: Is holding people morally responsible really the ideal way to interact with others?

The very idea of moral responsibility, he suggests, seems to bring about negative emotions toward other beings—emotions like hatred, anger, animosity—which are exactly the kind of feelings the Buddhist teaching advises us to abandon in order to become liberated. Then why can not Buddhist, or all of us, just embrace the fact that there are no such thing as morally responsible agents and liberate ourselves from the suffering of harboring agonizing hatred against others, which will also likely result in overall less pain and suffering since a possible repetition of retributions would be prevented? He cites various passages from Buddhist texts that offer a way to abandon hatred or anger by resorting to a typical Buddhist analysis of no-self: You are angry at someone, but "what is it you are angry with? Is it head hairs you are angry with? or body hairs? or nails? [……]"

A hard determinist analysis of agents' actions helps us to achieve the same kind of insight, he suggests. If this act of beating performed on me was something caused by factors that have nothing to do with this person, why should I be angry at the person? If this person is just doing what was forced upon him by the law of nature and prior state of the world, like an avalanche that struck my house, why should I harbor resentment against him? And indeed, we see from Śāntideva's writing the exact same advice about latent anger toward a person who harmed me. "Why am I angry with sentient beings? They too have causes for their anger." To sum up, Goodman argues that hard determinism is true, and that there is nothing to worry about it.

Although Goodman's argument presents a new perspective for understanding ethical implications of determinism, it has one major weakness, that is, Buddhists were compatibilists.

Resorting to the authority of the scriptures is not a sound way to make a philosophical argument, but if we are trying to propose a "Buddhist" theory of determinism and free will, it necessarily should be consistent with the Buddhist scriptures. We see in the Pali Nikāya the following passage:

> Having approached the priests & contemplatives who hold tha t…… 'Whatever a person experiences…… is all caused by what was done in the past,' I said to them: 'Is it true that you hold tha t……"Whatever a person experiences…… is all caused by what was done in the past?"' Thus asked by me, they admitted, 'Yes.' Then I said to them, 'Then in that case, a person is a killer of living beings because of what was done in the past. A person is a thief…… unchaste…… a liar…… a divisive speaker…… a harsh speaker…… an idle chatterer…… greedy…… malicious…… a holder of wrong views because of what was done in the past.' When one falls back on what was done in the past as being

> essential, monks, there is no desire, no effort [at the thought], 'This should be done. This shouldn't be done.' When one can't pin down as a truth or reality what should & shouldn't be done, one dwells bewildered & unprotected. One cannot righteously refer to oneself as a contemplative. This was my first righteous refutation of those priests & contemplatives who hold to such teachings, such views.[1)]

Here, the Buddha is explicitly criticizing the hard determinist view that everything a person does is a result of prior causes and thus a person does not have moral responsibility or free will ("desire"). If someone holds this view, then that person is a "killer of living beings," because it would render all living beings equivalent to objects lacking life or desires, like a stone, a chair, a car, or a robot. Moreover, if hard determinism is true, then we will not be able to discern "what should and what shouldn't be done," because any sentient being's moral responsibility would be negated. This argument raises a strong challenge to Goodman's view directly: If everything I do is determined by prior causes so that I am not responsible for anything I do, nor really have a choice about what I do, why should I rather take this peaceful approach toward a person who harmed me than just be angry at that person?

If so, were Buddhists indeterminists? It seems quite obvious that in this passage the Buddha is rejecting the idea that whatever a person does is determined by prior causes. However, in the same sutra, the Buddha goes on to deny indeterminism that things arise "without cause, without condition" to have the exact same implication of hard determinism: denial of life, will, and moral responsibility. Because if there is no causal connection between an action of a person and a will of a person, then consequently that person can not be held responsible for his action, because he never

---

1) *Tittha Sutta, Access to Insight.*

"caused" that action to happen.

Then does this mean that early Buddhists embraced an agent who by virtue of initiating a causal series can ultimately claim to be morally responsible for what he does? But as I have mentioned before, agent causation theory also implies denial of determinism. If an agent can be the ultimate starting point of a causal series, then this implies that at least some events can occur without any prior cause. Thus agent causation conflicts with Buddhist law of causal dependence, which states that everything has causes for its existence.

It seems that we have returned to the same dead-lock here. Buddhists definitely hold determinism to be true, but as we have seen, also dismiss its logical consequence that people do not have free will and moral responsibility. What should we make of this claim?

Buddhist Paleocompatibilism

In the same *sutta*, the Buddha, after denying both hard determinism and indeterminism, chooses to explicate his teaching of dependent origination and cessation once again. This might seem to contradict his first argument against the view that everything a person does is the product of prior causes, but it is not so. Because here, even though he is proposing a strictly determinist explanation, the word "person" is not used. What is used instead are mental and physical elements that constitute a person: ignorance, volition, consciousness, six faculties, contact, sensation, craving and so on.

I think this whole *sutta* supports Siderits's Buddhist paleocompatibilism, argued in "Beyond Compatibilism: A Buddhist Approach to Freedom and Determinism" (1987) and "Buddhist Paleocompatibilism" (2010).

Buddhist paleocompatibilism derives its name from classical compatibilism which asserted that it is "persons" not "wills" that can be considered to be free or not. Likewise, Siderits reminds us that it is only persons that can be either free or not free, either morally responsible or morally not-responsible. To Buddhist reductionists this paleocompatibilism implies that moral responsibility or free will is only a matter that can be discussed in the discourse of conventional level where we employ fictional concepts like "persons" to conveniently designate ultimately real atoms arranged in a certain way. In the ultimate level where there are only indivisible parts, the atoms, there can not be anything morally responsible or morally not-responsible.

Thus for example the Consequence Argument, as I have pointed out previously, is violating "two-way semantic insulation" that you can not talk about the whole and its parts at the same time, in the same sentence; by talking about causal series of ultimately real atoms together with conventional concept persons; more precisely, by confusing substance causation and event causation, the former requiring wholes made of parts and the latter only parts. Why this semantic insulation is more than quasi-philosophical dogma can be further argued. Siderits suggests the sorites puzzles as an example. Once we accept sentences that speak of both ultimate things and conventional beings like "'this cup' consists of such-and-such 'atoms'" as a valid sentence that has a truth value, then we will eventually have to abandon the principle of excluded middle. All we can say is that "what is conventionally denoted with the concept cup is a whole made of such and such atoms." So Siderits argues that, "There is no sound argument for the incompatibility of determinism and moral responsibility," and therefore, "In the absence of any compelling incompatibilist argument, we should accept the common-sense view that persons are generally morally responsible for their actions."

Against Siderits's Buddhist compatibilist understanding, Goodman

asks, “But doesn’t the view he advocates lead to the following conclusion: that, at the level of ultimate truth, nothing in the universe has what it takes to be a free agent? And isn’t this just a way of denying free will?” Yes, but no. “In the ultimate level, nothing in the universe can be morally responsible” is a true sentence for Buddhist paleocompatibilist, but “In the ultimate level, nothing can be morally not-responsible” is true as well. So in the ultimate level, they don’t just deny free will, they neither affirm nor deny free will. Because talking about free will in the ultimate level would by itself be a categorical mistake, a broken speech, without any “truth value”, or if they had one, it would be false. This is a similar case with the famous sentence “the King of France is bald.” When I point out there is no such thing as the King of France in this world, I do not thereby argue that therefore there does not exist hair of the King of France after all, and thus he is bald. I just mean that there is no King of France to be either hair-less or hairy. What Buddhist paleocompatibilist is suggesting by denying the application of persons in the ultimate level discourse is also the same. Sentences violating semantic insulation are either false, or lack enough seriousness to be either true or false.

The Buddha in the previous *sutta* dismisses both hard determinist’s position that what “a person” does is all determined by prior events and also indeterminist’s position that what “a person” does is all with out causes, but at the conclusion reaffirms the law of dependent origination. This seemingly mysterious denial of both determinism and indeterminism, and succeeding reaffirmation of determinism can be understood as prevention of violation of semantic insulation that you can not speak of a conventional concept “a person” together with a sentence of the ultimate discourse that “every occurrence of atoms has prior causes.”

So we can say Buddhists affirm the sentence “A person can be

the final traceable cause of an event and can be responsible in that case" and also affirm the sentence "Every occurrence of atoms has prior causes" by virtue of two-way semantic insulation, the first sentence being true in the conventional discourse, and the second in the ultimate discourse. Whereas causal analysis of an action performed on me can indeed prevent us from being possessed by negative emotions like resentment, and should in appropriate occasions be advised to see our situations thus, this approach can not exhaust the Buddhist position on free will as the scripture holds to the existence of moral responsibility in the conventional level, nor should it because unconditionally tracing a person's actions into the causal series of impersonal parts would result in denying any imperative for morally right behavior, including dispassionate forbearance Goodman is suggesting.

However, Siderits seems to hesitate from giving an affirmative account about what we conventionally speak with the expression such as "this person lied, killed, assaulted out of his free will, and is morally responsible for those actions" would correspond to in the ultimate level. I think his reluctance lies at the understanding that if we attempt to seek for a causal explanation of a person's action in the ultimate level, if we "pop the hood", to use his expression, we will not be able to find any absolute, final cause of that action. Thus he says,

> One way of bringing out the difficulty here is to ask the Buddhist Reductionist why it is that if strictly speaking there are no persons, it should be useful to hold that there are, and attribute moral responsibility to them. The response will be consequentialist in nature: our institutions of moral praise and blame help maximize overall utility, and these institutions require that we think of ourselves as persons.

However, although a sentence being conducive to less overall pain and suffering might be a sufficient, and only, condition for that sentence to be a conventional truth for Mādhyamikas, I don't think it is also the case with the Buddhist reductionists. Whereas Mādhyamikas would consider the sentence "here is a cup" to be conventionally true if and only if acting on the supposition that this sentence is true brings about less pain (less discomfort from thirst; given also that using this cup will not result in more pains in the future because for example I had to steal this cup in order to use it, or because I am developing a substantialist theory that this cup is something really real), Buddhist reductionists would consider the same sentence conventionally true if and only if there really exist in the ultimate level hardness atoms and whiteness atoms in a certain arrangement, efficacy of this sentence only being a byproduct of its corresponding relation to the ultimately real things. Thus for Buddhist reductionist there must be a distinctive arrangement of ultimately real atoms for the sentence "John killed Smith out of his free will" or "John is morally responsible for the death of Smith," and another kind of distinctive arrangement of ultimately real atoms for the sentence "Although John killed Smith, he is not morally responsible for his death." Taking Mādhyamika theory of truths is not really an option here because they don't think the law of dependent origination to be ultimately true.

Moreover, if Buddhist paleocompatibilism fails to give a different, distinct explanation for morally responsible actions in the ultimate level, it will still be subject to the Consequence Argument. Buddhist paleocompatibilist's objection to the Consequence Argument as presented in this paper was that it would be wrong to derive (3) from (1) and (2), (1) being a sentence in the ultimate level discourse and (2) being a sentence in the conventional level discourse. But we can make a little adjustment to the Consequent Argument and argue

that people do not have moral responsibility in the conventional level as well. Let us assume that J did A at *t*.

(1) J is morally responsible for doing A, only if J could have done not-A.
(2) What is conventionally designated with the expression "J is doing A at *t*" and "J is not doing A at *t*" correspond to different arrangements of atoms in the ultimate level.
(3) In the ultimate level, determinism is true.
(4) If determinism is true, the occurrence of arrangement of atoms at *t* we conventionally denote with the expression "J is doing A" is determined by prior states of the atoms and the laws of nature.
(5) Thus, there being at *t* a different arrangement of atoms that we would conventionally designate with the expression "J is not doing A at *t*" rather than "J is doing A at *t*" would require either a different laws of nature or a different states of atoms in the time before *t*.
(6) Either the laws of nature or states of atoms in the past can not be altered by definition.
(7) Thus, an arrangement of atoms that can be conventionally designate with the expression "J is not doing A at *t*" is impossible to exist.
(8) Therefore, J could not have done not-A at *t* and does not have moral responsibility for doing A.

In short, determinism requires that only one possibility exists at one moment, namely, what the state of atoms in the world is at that moment. Therefore, there can not be alternate possibilities in the ultimate level. If there are no alternate possibilities in the ultimate level, then it is false to postulate a conventional sentence that requires alternate arrangement of atoms in the ultimate level. If so, "J could have done otherwise than A" is a conventionally false

sentence as how “there is a cup on my desk” is a conventionally false sentence when there are no certain arrangement of witness atoms and hardness atoms in that space. And if we accept the Principle of Alternate Possibilities to be true, than there can not be moral responsibility or free will even in the conventional level discourse.

These are two challenges I think Siderits’ Buddhist paleocompatibilism is facing. But Frankfurt’s argument for falsity of the Principle of Alternate Possibilities and his analysis of morally responsible action can provide Buddhist paleocompatibilists a way out.

Frankfurt-style Analysis of Morally Responsible Action of an Agent

In “Alternate Possibilities and Moral Responsibility” (1969), Frankfurt suggests a thought experiment later known as the “nefarious neurosurgeon case” or “Franfurt-style case” which refuted the Principle of Alternate Possibilities that one must have alternate possibilities in order to be held morally responsible for one’s action. I am going to briefly reconstruct his argument with some alterations for my own use.

Suppose that John is a locomotive driver. While driving his locomotive around the route, he sees in a distance an old, powerless man sitting on a wheelchair that seems to have stuck tight in the rail. From the old man’s facial expression John knows that this man is alive and that he wants to stay so. But John, after having a depressing day, which was started by finding out that his wife left him and the news from his boss that his payment was cut down for repetitively being late to work, instead of hitting the brake, develops an urge to just speed up and run over the old man and decides to carry out that urge, and does it indeed, causing the old

man's death. But in that day earlier in the morning, John's colleague Smith did not pay enough attention when maintaining the locomotive that by the time when John hit the old man, the brake was totally useless.

In this case, no one would disagree that John was the one who is responsible for the old man's death and that he killed the old man out of his free will, despite the fact that he could not have done otherwise. So PAP is false.

In a consequent essay "Freedom of the Will and the Concept of a Person" (1971), Frankfurt further goes on to provide an alternative definition of a morally responsible act of a person. A person is morally responsible for what he does if and only if there is accordance between his action, his volition, and his meta-volition.

I think what Frankfurt-style cases and his definition of morally responsible action are telling us is that PAP is a misconstrued principle stemming from superficial comprehension of our conventional usage of the argument that "if one could not have done otherwise, then one is not responsible for that action." PAP is basically contraposition of this argument that, "one is responsible for one's action, only if one could have done otherwise." But in the previous conditional, it is already assumed that he tried to do otherwise, that is, he developed an urge to do otherwise, or at least speculated on the possibility of doing otherwise, but that he could not, which then without any relation to alternate possibilities just means that there was a discordance between his volition, meta-volition and action. This is what we mean when we tell our boss after being late to work that "I am late, but I could not have done otherwise." Things just happened against my volition that there was nothing I could have done about it, that there was discordance between my will and my action. But superficial contraposition of this conditional has a very different implication, that a person in some sense must

be capable of creating alternate courses of universe to be morally responsible, which is, of course, not compatible with determinism.

In the example of John, he is morally responsible for the old man's death despite the fact that he had no alternate possibilities, because he had the volition to kill that man, and meta-volition to grant that volition, and then physical action of carrying out that volition. And even though there are at least five people who provided causes for this unfortunate happening — the old man, John's boss, John, John's wife, and probably the old man's nurse, provided that she was not at the scene in search for help — we would without any hesitation say that it was John rather than any others who is responsible for the death, because unlike John, they did not have the volition to kill that old man when they provided those causes that eventually took part in the old man's death, that is, they lack the accordance of volition and physical action.

The fact that development of John's urge to kill that old man can be further traced back to persons and situations around John is irrelevant for finding a morally responsible person, unlike how many people worry. Once we have reached a psycho-physical causal series where the accordance of volition, meta-volition, and action that brought about relevant event was first formulated, we conventionally say that this is the person responsible for the action. Manipulation objection to Frankfurtian interpretation of freedom is wrong because it presupposes prior occurrence of accordance between volition, meta-volition, and action that caused the event. If there was a nefarious neurosurgeon who even created that very volition of John in the first place, accordance of volition, meta-volition, action can be further traced back to that surgeon, getting John off the hook.

For Buddhist reductionists Frankfurt's analysis provides a way to explain what we conventionally denote with the expression "John is responsible for the man's death" and "John's wife is not responsible

for the man's death" in the ultimate level. That is, what we conventionally designate with the phrase "a morally responsible action of A" or "an action done of A's own free will" is a whole made of parts. Primary parts being the following three causal relata: a volition that belongs to the causal series we conventionally denote as a person A, a meta-volition that belongs to the causal series we conventionally denote as a person A, and a physical action that belongs to the causal series we conventionally denote as a person A.