Human Nature and Moral Responsibility

Cameron Davis
Johns Hopkins University

International Undergraduate Philosophy Conference

Montana State University, Bozeman

September 6-7, 2013

Preamble

Holding others responsible for and responding resentfully to their wrongdoings are nearly universal practices. Few philosophers and social activists have ever adopted the idea that one should, seemingly against his nature as a human being, seek to completely abandon his "negative reactive attitudes", as P.F. Strawson coined them in *Freedom and Resentment.* The notion that one should suspend all negative reactive attitudes such as anger and resentment is based on the idea of determinism: that all events, choices, and actions are causally determined and cannot happen in any other way—every event is predestined and must occur as part of an immutable sequence of events. Strawson does not refer to a specific type of determinism, but perhaps an explanation of the reasoning behind the main conception of determinism will be helpful.

The main conception of determinism, physical determinism, is based on the important premise that all events are physical events, and even mental events such as thoughts and decisions are simply physical events or the results of physical events in the brain. If all events are physical events or the results thereof, then all events are causally constrained to be as they are—they could not be otherwise—because they are simply parts of a sequence of events operating under universal, invariable physical laws such as energy, mass, etc. Our mental events, which according to physical determinism are physical events, are no different—each thought and decision is not a result of someone's free will but rather is one neural event (such as a pattern of neurofiring) in a sequence of neural interactions. Many understandably hold that if determinism is true, humans have no free will and are simply characters in a prewritten story of causally related physical events. While there is a sect of determinists known as "compatibilist determinists," who believe that determinism and free will are compatible, for the sake of

simplicity, in this essay, the terms "determinism" and "determinist" will refer to a conception of determinism that is incompatible with free will.

In today's society, where free will pervades the worldview and determinism is widely considered a high-flown philosophical fantasy, negative reactive attitudes are expressed in the vast majority of cases in which a moral actor commits a wrongdoing—but not all. If an agent harms another only accidentally and does not express ill will, or if a sociopath or someone with a neurological abnormality has no meaningful control over his behavior, he is often met with forgiveness rather than anger or disapprobation. Even in our free-will dominated society, many still find reasonable justification for suspending negative reactive attitudes some of the time, but if determinism is true, it seems no one is ever responsible for his behavior. If determinism is true, one's will is causally determined, simply a product of the interactions between the physical events that preceded it and the physical laws operating on those physical events, and one has no meaningful control over it. No one chooses in any meaningful sense to have their will but simply receives it as a causally determined inevitability. Strawson, in agreement with most determinists, asserts that if determinism is true, the rational practice is to suspend **all** negative reactive attitudes, since it is nonsensical and unfair to hold others responsible for actions they have no control over. However, he argues that **despite the rationality of abandoning these attitudes, humans should not attempt such a feat, as it is surely so antithetical to human nature as to be impossible.**

Freedom and Resentment has gained significant popularity as one of the most influential philosophical papers of the late 20th century. Strawson's idea is intuitively appealing because it accords with societal norms. Negative reactive attitudes such as anger and resentment toward wrongdoing are so common and so widely accepted as justified that these attitudes seem to be

insuppressible parts of human nature. In this paper, however, I will argue that this is not necessarily the case; humans can control and suspend their negative reactive attitudes in the appropriate cases, including in the case that determinism is true. In arguing that humans are capable of suspending all of their reactive attitudes based on belief in determinism, I will show that Strawson inadvertently substantiates my claim.

Strawson's argument is self-defeating. He admits that under certain circumstances, one might understand why an actor committed a wrongdoing and consequently suspend reactive attitudes, taking an "objective" view. One of these exempting circumstances, according to Strawson, is "derangement" or "perversion." By this very admission, Strawson undermines his own argument that humans could not suspend negative reactive attitudes based on belief in determinism. He does so by establishing that humans can control their reactive attitudes by reasoning about whether the actor is responsible for his will. I propose that if people can reason that a moral agent committed a wrongdoing because he was "deranged" or "perverted" and can consequently suspend negative reactive attitudes, as Strawson admits, they should be able to perform a similar suspension by reasoning that a moral agent committed a wrongdoing simply because his ill will and behavior were causally determined. This paper has three main sections. In the first, I will reconstruct Strawson's argument. In the second, I will elucidate its logical incoherency. In the third section, I will examine Strawson's resources to counter and then refute these possible counterarguments.

Section I: Reconstructing Strawson's Argument

First, I will reconstruct Strawson's argument in more detail. Determinism, i.e. an incompatibilist conception that rejects free will, is an extremely controversial theory with little support even from the philosophical community, but if it is true, it might render nonsensical and

unfair society's ubiquitous practice of holding people responsible for their actions. If determinism is true, one has no control over his actions or will, so how can one be held responsible for them? Strawson took a unique approach to addressing the implications of determinism on moral responsibility. He argued that reactive attitudes such as anger and resentment toward someone who manifests an ill will are rooted in human nature. Contrary to the common conception of moral responsibility, Strawson believes that these attitudes are prior to and constitutive of moral responsibility. People do not react angrily or resentfully to ill will because they hold people morally responsible. Rather, Strawson believes, moral responsibility—the holding of others responsible—results naturally from people's insuppressible psychological tendencies to react angrily and resentfully toward ill will, so the truth of determinism is irrelevant to the issue of moral responsibility.

Strawson then admits that people refrain from reacting in such a way in some cases of exemption, and it is with this admission that such exemptions are commonly made that Strawson undermines his argument. He presents two main types of exempting cases. The first type is what I will call "exemptions by circumstance." Sometimes a person may do something that would naturally cause another anger, such as shove him, but not out of ill will—not with desire to harm. Imagine a case in which the person had to shove another in order to reach and rescue a dying loved one or a case in which the shove was done by accident—perhaps one's shoe came untied, he tripped, and the shove was a byproduct of his accidental fall. Strawson explains that in these cases one may suspend reactive attitudes by reasoning that the other person did not display ill will. While these cases are interesting and perhaps they could play some role in clarifying the incoherency in Strawson's argument, cases of the second type are much more elucidating.

I will describe the second type of exempting cases as "exemption by inapplicability." Unlike the previous type, the wrongdoer intends to harm, expressing the ill will that Strawson believes is essential to eliciting negative reactive attitudes; however, people generally can suspend the attitudes, Strawson admits. He does not explain why, though. He simply admits to this general ability to suspend attitudes toward certain types of people. Strawson's examples include children and, what I will focus on, a "deranged" or "systematically perverted" person. People tend to refrain from holding morally responsible actors of certain circumstances—e.g. those with neurological defects or sometimes those that have suffered exceptionally hardening pasts—when they find the actors' behavior can be explained by their "unfortunate formative circumstances," in Strawson's words. While discussing a murderer, one tends to react with anger, disgust, and vicarious resentment. On the other hand, Strawson postulates that if one knows a murderer to be a victim of a neurological disorder that has rendered him unable to empathize or uncontrollably violent and malicious, one tends not to react with anger or vicarious resentment, but rather to take an "objective attitude": to see that person as a subject of policy, someone who must be pacified, softened, and possibly controlled. Herein Strawson's admission undermines his own argument.

Section II: Identifying the Logical Incoherency

Strawson's examples clearly illustrate that humans have the capacity to reason about the reactive attitudes they should and should not have. He himself has proven that they are not the insuppressible emotions that he describes them to be. To see the inconsistency more closely, we must answer the questions Strawson tellingly fails to: What unites the cases of "exemption by inapplicability," and why can we suspend negative reactive attitudes for deranged, perverted, or neurologically abnormal people? The answer, which I hope the reader will find intuitively

appealing, is that **one tends to reason the neurologically abnormal or deranged actor is not responsible for his ill will**. This interpretation makes sense in light of how society tends to think about these types of people. Neurologically abnormal people who commit atrocities are often thought of as unable to control their intentions and behaviors—i.e. their wills. Similarly, the "deranged" or "perverted" are often considered to have wills shaped into depravity by their "unfortunate formative circumstances." In both cases, I believe most would agree with the proposition that they are able to suspend reactive attitudes toward these actors because they find the actors are **not responsible for their wills and thus not responsible for their ill will.**

If this answer is correct, Strawson's argument seems to fall apart. If one can suspend attitudes on account of a deranged or neurologically abnormal person's not being responsible for his ill will, why can't one do the same for all given a belief in determinism? **In a deterministic world, no one is ever responsible for his ill will.** Perhaps Strawson is right that there is something natural in feeling anger and resentment toward those with ill will, but since humans are capable of taking the unemotional, rational, "objective" stance toward deranged or perverted people, perhaps what is human nature and maybe even insuppressible is the tendency to react with anger and resentment **only when one believes the actor is truly responsible for his ill will**. **If determinism is true and humans lack free will, no one is responsible for his ill will**, and as Gandhi and Albert Einstein espoused, it makes sense to desire and demand fair treatment but not to react angrily, resentfully, and certainly not vengefully toward one's mere annoyances or even oppressors.

Section III: Examining A Possible Reply

Strawson could reply by arguing that humans lack the willpower or endurance to suspend all negative reactive attitudes. They can do so in rare, extreme cases, but it is in their nature to

express at least some negative reactive attitudes toward their plights and the injustices in the world. This argument faces some major problems. First, it seems wrong on a very basic level. When humans realize that wrongdoers are out of control of their actions, they do not seem to expend some of a putatively finite store of willpower in suspending negative reactive attitudes. Generally, it is not a struggle to suspend negative reactive attitudes toward a person who is known to have been born neurologically abnormal. Consider a person with Tourette's syndrome who might unwillingly shout profanity during a sacred ceremony. One somewhat easily suspends his negative reactive attitudes. It does not require depletion of willpower; it simply requires a realization that the actor is not in control of his actions or that the actor does not have an ill will, so there is no reason to be mad or resentful toward him. If willpower to suspend negative reactive attitudes is not limited, and suspension simply requires a realization that one is out of control of his actions, then suspension of these attitudes toward all based on determinism seems entirely possible. The second major problem with this argument now directly follows from the example of a man with Tourette's. Suspension of negative reactive attitudes would not deprive humanity of all negative emotional reactions, which seem to serve a cathartic role in helping people to cope with plights and injustices. The negative emotional reactions would simply change. Instead of feeling angry and resentful toward an actor one finds deserving of disapprobation, repugnance and vengeance, one might feel sad, disappointed, or frustrated by simply unfortunate events. This is probably how most would react to the aforementioned case of a person with Tourette's shouting profanity during a sacred ceremony. Finally, since Gandhi, Albert Einstein, and presumably many determinists have been able to suspend negative reactive attitudes on a broad scale, this argument is empirically denied as well.

Strawson's mistake is understandable. Practices of reacting resentfully and holding others responsible for their actions are so nearly universal that, prima facie, such practices seem to be human nature. However, Strawson's argument may be limited by its shortsightedness. Strawson believed negative reactive attitudes are insuppressible and part of human nature simply because they are so common, but perhaps they are only so common because the existence of free will and responsibility for one's actions are so widely and wholeheartedly accepted as truth, so engrained in society's worldview. Given that people can indeed think rationally about whether it makes sense to have negative reactive attitudes toward a "deranged" or "perverted" person, it is likely that widespread belief in determinism would produce a society of people who take the "objective" stance toward all unfortunate circumstances around them. Strawson seems to have been limited by his shortsightedness. It does not seem that humans in general can abandon negative reactive attitudes toward everyone now, as most people scoff at such a notion, but with a widespread belief that everyone's actions are entirely out of his or her control, subject to the forces of determinism, a near universal suspension of these attitudes should be considered possible. That is, in a world where a belief in free will is not so pervasive and engrained, suspension of reactive attitudes may be much more natural.

Recent Empirical Inquiry

Strawson, if he were alive today, might believe he has gained some support from evolutionary moral psychology. The idea that moral condemnation and moral emotions such as anger and resentment are rooted in human evolutionary history has recently gained significant empirical support. Some evolutionary psychologists have proposed theories of "indirect reciprocity", which state that the human brain's pleasure centers may adaptively reward certain moral behaviors because others notice such moral behaviors and are willing to form cooperative

and mutually beneficial relationships. Similarly, some evolutionary psychologists have argued that the brain may be programmed to condemn and react angrily and resentfully toward immoral behaviors because such reactions also broadcast a benevolent disposition and willingness to form cooperative, mutually beneficial relationships. If negative reactive attitudes are rooted in human evolution, they may indeed be part of human nature, as Strawson believed.

However, there are two main problems with the jump from these recent findings to the conclusion that Strawson was right that humans could not suspend all reactive attitudes based on a belief in determinism. First, even if negative reactive attitudes are evolutionary mechanisms, they are not necessarily insuppressible. Just because something is some way in its natural state does not mean that it has to be so. Secondly, negative reactive attitudes clearly are not insuppressible. The exempting cases that Strawson provides show that to be the case. If one believes that a moral actor is not responsible for his ill will because he is neurologically abnormal, one can overcome his evolutionary tendency to react with anger and resentment, and similarly, with a belief in determinism, he should be able to overcome this evolutionary tendency for negative reactive attitudes toward people in a deterministic world who are not any more responsible for their ill wills than they are for the chain of causally related physical events that surround and affect them.

Summary

Strawson's *Freedom and Resentment* is an extremely influential work on determinism and moral responsibility. His idea that, even if determinism is true, suspension of negative reactive attitudes for all is so antithetical to human nature as to be impossible is a unique take on the issue. However, Strawson's argument exhibits a major logical inconsistency that undermines his entire argument. He willingly admits that people can suspend their negative reactive attitudes

if, with rational consideration, they find good reason to. With this admission, he incites the reader to ask why this same sort of rational consideration and consequent suspension of attitudes and responsibility is not possible with regard to all human action. Even if people are evolutionarily designed to react angrily or resentfully toward ill will, since they can suspend their negative reactive attitudes toward deranged or neurologically abnormal people who cannot control their wills, they should be able to do the same toward everyone given they believe in determinism, for anyone in a deterministic world has even less control over his will than a deranged, neurologically abnormal person in a world with free will. These attitudes may seem natural and insuppressible simply because the idea of free will is so engrained in society's worldview, but if the common man saw the world as a determinist does, perhaps anger and resentment would become the unnatural reactions.

Strawson's greatest mistake was setting for himself an unattainable burden of proof—one that he failed to come close to meeting. *Freedom and Resentment* was meant to end the debate between the "optimists" and the "pessimists". The optimists believe that even if no one has free will, society should continue to hold people morally responsible to regulate behavior. Pessimists, on the other hand, believe that holding others morally responsible despite their lacking free will is unfair and unjust. Strawson believed his paper ended this debate because it showed it doesn't matter whether it's right or wrong to hold morally responsible people without free will—we can't help but do it. To meet this burden—to end the debate between the optimists and the pessimists—Strawson needed to prove that **no one** could **ever** suspend negative reactive attitudes given a belief in determinism. Even if you believe not everyone could suspend attitudes in every case, it should be clear—in light of the logical inconsistency and cases like Gandhi and Einstein—that at least some people could do it some of the time. Given this, the debate between

the optimists and pessimists remains valuable. In a world without free will, is it better to uphold moral responsibility in the name of behavioral regulation or to abandon it in the name of fairness? Strawson set out to end this debate, but it remains just as important for the philosophy community to confront as ever.

REFERENCES

Alcock, John. *Animal Behavior*. 9th ed. Sinauer Associates, Inc., 2009. Print.

Strawson, Peter. "Freedom and Resentment." *Perspectives on Moral Responsibility*. (1963): n. page. Web. 1 Jun. 2013.