

# Agent Causation and the Alleged Impossibility of Rational Free Action<sup>1</sup>

---

Chris Tucker  
Purdue University  
[tuckercs@purdue.edu](mailto:tuckercs@purdue.edu)

*I think the impossibility of free will and ultimate moral responsibility can be proved with complete certainty.*<sup>2</sup>  
Galen Strawson

This statement is rather bold, but I must disagree with Professor Strawson. One argument, perhaps *the* argument, which Strawson believes to prove the impossibility of free will and moral responsibility is my focus in this paper. This argument takes the form of an infinite regress which supposedly shows that rational actions can't be free. In section II, I hope to present Strawson's argument as clearly and as simply as possible. This will require deviating from the way that Strawson himself summarizes the argument.<sup>3</sup> My ultimate aim in this paper is to show that agent causation theorists need not be troubled by Strawson's famous argument.

## I. An Introduction to Strawson's Lingo

In order to understand the regress, we will need to introduce four Strawsonian terms and two of his principles. It is imperative that we pay careful attention to these terms, because their precise meaning may be subtly, but importantly different than what we might have expected.

The first term we should consider is:

**Rational Action:**  $S$ 's action  $A$  is rational only if it is fully determined by  $S$ 's reasons  $R$ .

---

<sup>1</sup> Word Count: 2,999. Thanks to Michael Bergmann for reading an earlier draft of this paper.

<sup>2</sup> Quoted from an interview with Strawson, conducted by the *Believer*, March 2003. Stable URL: [http://www.believmag.com/issues/march\\_2003/strawson.php](http://www.believmag.com/issues/march_2003/strawson.php).

<sup>3</sup> See pages 28-9. Unless otherwise noted, all page number refer Strawson's *Freedom and Belief*. Two additional notes are worth mentioning. First, Strawson's 1994 (6-7) and his 1995 (15-6) present very similar summaries. Second, I should stress that my hope in deviating from his presentation is to present the argument better than he did. I will often notify the reader of deviations and will justify those deviations in the footnotes.

We should be careful to appreciate that this sense of rational action has no normative component; says Strawson, “In this sense one may perform a rational action even if one’s reasons for that action are highly irrational from some ordinary point of view” (28, footnote 5). This notion of rational refers to something purely psychological.<sup>4</sup>

Notice that rational action is defined in terms of full determination. Thus, we should turn our attention to that concept.

**Full determination:** *S*’s reasons *R* fully determine his action *A* only if the rational explanation of *A* that cites *R* and only *R* is a true, full explanation of *A*.<sup>5</sup>

Again, it will be helpful to see what is *not* being claimed when we say that reasons *fully determine* actions. It is not claimed either that reasons *cause* the action, or that they determine it in the sense of necessitating it, or making it such that the agent *had* to act in the way he did. According to Strawson, when he speaks of ‘fully determinative,’ he has the following sort of connections in mind:

*a* [i.e. the agent] turns on the light, because he hears something moving in the room, wants to see what it is, and believes that the way to satisfy this want is to turn on the light—and *that’s all there is to it*. *a*’s having the desire and belief fully explains his action so far as it is a turning on of the light, and is to the same extent fully determinative of it. (38)

Hence, even if the agent—rather than the reasons—causes the action, as long as the agent performs the action for those reasons, or acts on those reasons, the reasons fully determine—at least in Strawson’s sense—the action.

---

<sup>4</sup> Some of O’Connor’s criticisms miss the mark, because he fails to appreciate that Strawson’s use of ‘rational’ has no normative component (1995b: 6-7).

<sup>5</sup> Here is the definition of full determination stated in Strawson’s own words: “*a*’s reasons *R* fully determine (are fully determinative of) his action *A* just in case the rational\* explanation of *A* that cites *R* and *R* only is a true and full rational explanation of *A*” (37). You will of course notice that Strawson makes a distinction between ‘rational’ and ‘rational\*.’ My use of ‘rational,’ however, is equivalent to his use of ‘rational\*.’ For Strawson, the difference is one of scope: “True full rational explanations differ from true full rational\* explanations simply in this: they may cite desires and beliefs that the agent has not got, as well as those it has got” (37).

As you can see in the above definition, “the notion of full determination is simply *defined* in terms of true full explanation” (37, emphasis original); therefore, the next term we need to introduce is full rational explanation.

**Full Rational Explanation:** A rational explanation is full only if it is one that, by citing *R*, everything is cited about *S*’s mental states that made it the case that he performed *A*.<sup>6</sup>

It is very important that we notice that to be a full rational explanation it is not required that *R* be *everything*, or even everything *about S*, that made it the case that *S* performed *A*. It is only required that *R* be everything about *S*’s *mental states* that made it the case that she performed *A*.

Finally, and most straightforwardly, we should define:

**Free Action:** *S*’s action *A* is free only if the *S* was responsible for *A*, that is, capable of being held morally accountable for *A*.<sup>7</sup>

Now that we have discussed all the terminology necessary for understanding the regress, it is time to turn our attention to two principles which drive the regress. Informally and very generally, Strawson’s two plausible principles are ‘if one is responsible for their actions, then she is responsible for the way she is’ and ‘if one is responsible for the way she is, then she must have rationally chosen to be that way.’ The more specific and careful versions of these principles—the versions needed for the regress—are as follows:

**PP1:** If *S*’s reasons *R* at *t* fully determine *S*’s action *A* and *S* is responsible for *A*, then *S* is responsible for *R* at *t*.

**PP2:** If *S* is responsible for *R* at *t*, then *S* rationally chose *R* at *t*.<sup>8</sup>

---

<sup>6</sup> In Strawson’s words, an explanation “gives, while citing only *a*’s reasons, a full account of what it was about *a*, mentally speaking, that made it the case that he performed the action that he did perform” (36). Notice that I replaced Strawson’s ‘the way *a* is, mentally speaking’ with ‘*S*’s mental states.’ I make this replacement because, to me at least, it clarifies exactly what is at issue; however, if you wish, you can simply replace ‘*S*’s mental states’ with ‘the way *S* is, mentally speaking.’

<sup>7</sup> Strawson asserts, “In this book the word ‘free’ will be used in what I call the ordinary, strong sense of the word. According to which to be a free agent is to be capable of being *truly responsible* for one’s actions” (1, emphasis original). And according to Strawson, to be truly responsible is “to be capable of being truly deserving of praise and blame for them” (1). (It should be noted that my use of ‘responsible’ is equivalent to Strawson’s ‘truly responsible.’)

<sup>8</sup> In Strawson’s 2002, he presents four versions of the “Basic Argument” and at least three of the four explicitly rely on something like PP1. Compare (2) with 1.2, 2.2, 3.2 from that paper. In that same paper, 3.6 is a different version of PP2.

Strawson, as far as I can tell, doesn't really argue directly for these principles. I take it that he believes that these principles are intuitively plausible, and so the burden of proof is on those who object to them.<sup>9</sup>

## II. The Vicious Regress: Why Rational Free Action is Impossible

Now that we have introduced all of the relevant Strawsonian terminology and principles, we can present the regress, which hopes to show that rational action cannot be free.

- (1) If  $S$ 's reasons  $R_0$  at  $t_0$  fully determine  $S$ 's action  $A_0$  and  $S$  is responsible for  $A_0$ , then  $S$  is responsible for  $R_0$  at  $t_0$ .<sup>10</sup> [(PP1)]
- (2)  $S$ 's reasons  $R_0$  at  $t_0$  fully determine  $A_0$  and  $S$  is responsible for  $A_0$ . [suppose for *reductio*]
- (3) Therefore,  $S$  is responsible for  $R_0$  at  $t_0$ . [(1),(2)]
- (4) If  $S$  is responsible for  $R_0$  at  $t_0$ , then  $S$  rationally chose  $R_0$  at  $t_1$ .<sup>11</sup> [(PP2)]
- (5) Therefore,  $S$  rationally chose  $R_0$  at  $t_1$ —that is,  $S$ 's reasons  $R_1$  at  $t_1$  fully determined  $S$ 's choice for  $R_0$ . [(3),(4)]

---

<sup>9</sup> In his 2002, 3.6 is a different version of PP2, and he says that he will not defend 3.6 because “it is evident on reflection” (446). In that same paper 1.2 is a different version of PP1, and he calls it “obvious” (443). He does promise, however, that he will defend 1.2 or one of its “close cousins” later in the paper. I am not positive about this, but I believe his promise is fulfilled in section 6. If so, then his defense really is a response to a Leibnizian view of freedom. In section IV, I will argue that response, whatever its success against the Leibnizian view, cannot be successfully applied to agent causation.

<sup>10</sup> To distinguish my numbers from those of Strawson (pgs 28-9), I will symbolize Strawson's (1) as (S1). (1) is roughly equivalent to (S3): “If, therefore, one is to be truly responsible for how one acts, one must be truly responsible for how one is, mentally speaking—in certain respects at least” (28). Later in the chapter it becomes very clear that the relevant aspects of how one is, mentally speaking, are the person's *reasons* for acting. This is so, because what is of concern is *rational action*, action performed for a reason (see 33-5, for Strawson's understanding of reasons).

<sup>11</sup> My (4) is roughly equivalent to (S4): “But to be truly responsible for how one is, mentally speaking, in certain respects, one must have chosen to be the way one is, mentally speaking, in certain respects. (It is not merely that one must have caused oneself to be the way one is, mentally speaking; that is not sufficient for true responsibility. One must have consciously and explicitly chosen to be the way one is, mentally speaking, in certain respects, at least, and one must have succeeded in bringing it about that one is that way.)” (28). In (S5), Strawson clarifies what it means to ‘consciously and explicitly’ choose. He says, “One cannot really be said to choose in a conscious, reasoned fashion, to be the way one is, mentally speaking, in any respect at all, unless one already exists, mentally speaking, already equipped with some principles of choice, ‘ $P_1$ ’—with preferences, values, pro-attitudes, ideals, whatever—in the light of which one chooses how to be” (29). Whatever these principles of choice may be, e.g. preferences, values, etc., they would seem to be a *reason*, i.e. a desire or belief, for making oneself have certain other desires and beliefs. Hence, Strawson is treating ‘rationally choosing  $R_1$ ’ as equivalent to or a subset of ‘ $R_1$  must be the result of some rational action.’ Choosing  $x$  for some reason isn't merely forming an intention to do  $x$  for some reason, it also requires that the intention be fulfilled, that one succeeds in bringing about her intention.

- (6) If  $S$ 's reasons  $R_{-1}$  at  $t_{-1}$  fully determine  $S$ 's choice for  $R_0$  and  $S$  is responsible for choosing  $R_0$ , then  $S$  is responsible for  $R_{-1}$  at  $t_{-1}$ .<sup>12</sup> [PP1]
- (7) Therefore,  $S$  is responsible for  $R_{-1}$  at  $t_{-1}$ . [(4),(5),(6)]
- (8) If  $S$  is responsible for  $R_{-1}$  at  $t_{-1}$ , then  $S$  rationally chose  $R_{-1}$  at  $t_{-2}$ .<sup>13</sup> [PP2]
- ...
- (9) And so on. Therefore,  $A_0$  (or any rational action) can't be free, because it would require the actual completion of an infinite regress of rational choices—an impossibility for beings like us.<sup>14</sup>

### III. Rejecting PP1: How an Agent Causation Theorist Can Avoid the Regress

Since the regress, as I have stated it, is clearly valid, we must reject at least one of the premises, and I think the agent causation theorist can reject (1) and (6), the premises which depend on PP1. The basic approach of the agent causation theorist, then, is to reject PP1. There are two mistaken reasons why we might believe this approach is implausible, and it will be helpful to consider those reasons up front.

First, this approach will look implausible if we confuse (PP1) with:

**(PP1\*)** If  $S$ 's reasons  $R_0$  (and the laws of nature) at  $t_0$  necessitate  $S$ 's action  $A_0$  and  $S$  is responsible for  $A_0$ , then  $S$  is responsible for  $R_0$  at  $t_0$

By “necessitated” I mean that  $S$ 's having those reasons (and the laws of nature being the way they are) *logically implies*  $A$ . If  $A$  is necessitated by  $R$ , then  $R$  explains why  $A$  *had to occur*, even if  $R$  isn't the cause of  $A$ . (PP1\*) seems obviously true, or at least it does for everyone who is fond of the improved versions of Van Inwagen's principle beta. The phrase “fully determine” may beguile us into confusing (PP1) with (PP1\*), but we must remember what Strawson means by that very important phrase. For reasons to fully determine an action, it is

<sup>12</sup> My (6) is roughly equivalent to (S6): “But then to be truly responsible on account of having chosen to be the way one is, mentally speaking, in certain respects, one must be truly responsible for one's having *these* principles of choice  $P_1$ ” (29, emphasis original).

<sup>13</sup> My (8) is roughly equivalent to (S7): “But for this [i.e.,  $S$ 's being responsible for  $R_{-1}$  at  $t_{-1}$ ] to be so one must have chosen them in a reasoned, conscious fashion” (29).

<sup>14</sup> My (9) is roughly equivalent to (S9): “And so on. True self-determination is logically impossible because it requires the actual completion of an infinite regress of choices of principles of choice” (29). To be ‘responsible’ for the way one is, one must be ‘truly self-determined’ (26), that is, one must have freely done something that resulted in the way one is.

not required that those reasons (and the laws of nature) necessitate or logically imply the action; in Strawson's parlance, 'full determination' is not tantamount to 'causal determination' or 'necessitation.' Reasons fully determine an action only if those reasons and only those reasons cite everything about *S*'s mental states which (actually) contribute to the performance of the action. Thus, it is perfectly possible for some set of reasons to *fully determine* an action *and* for something else to also causally contribute to the production of the action. According to Strawson, reasons can fully determine our actions, even if the agent actually causes the action (thus being an additional causal contributor) because those reasons *fully explain* why the agent acted as he did.<sup>15</sup> After Strawson argues for this very point, he summarizes his position by arguing:

It can be true both (i) that *a* [i.e. the agent] and only *a* performed A [i.e. the action], and (ii) that to cite *a*'s reasons is to give a true and full explanation of A—although (ii) entails (iii) that *a*'s reasons were fully determinative of A in the present sense. For (iii) is entirely compatible with (i). (39)

And isn't it rather commonsensical to say that the *agent*, in addition to the reasons, contributes to the performance of the action?

This commonsensical possibility is precisely what agent-causation theorists have in mind. They suggest that the agent, while acting for reasons, must contribute causally by deciding to act on those reasons; the agent, while acting for reasons, is in *control* of which reasons she acts on. The agent causation theorist can say that the agent turned the light on because she wants to see what made some noise and believed that turning the light on would help her do that "*and that's all there is to it*" (38, emphasis original). (Yes, *a* can always do otherwise—insofar as *a* is free. Despite having actually turned the light on, *a* could have ignored the sound and not

---

<sup>15</sup> Strawson asks, "If we choose to characterize the agent as the *cause* or determiner of A, the next question is then this: *why* did *a* perform A (at time *t*)?... The answer is, of course, by reference to his belief-desire states [i.e. his reasons]" (39, emphasis on 'cause' is mine).

turned the light on; however, this does not disqualify the above explanation from being a good one or from being a full rational explanation.) Supposing that we do exhibit this kind of active control over both what we do and why we do it, we would seem to be free—free in the sense in which we would be morally accountable for our actions. And we can have this freedom even if we are not responsible for the reasons on which we act.

Second, if we are not careful, we may mistakenly believe that rejecting (PP1) commits us to rejecting one of the following principles:

- (a) If  $S$  is responsible for  $R$  and  $R$  fully determines  $A$ , then  $S$  is responsible for  $A$ .
- (b) If  $S$  is responsible for  $R$  and  $R$  necessitates  $A$ , then  $S$  is responsible for  $A$ .

To me at least, (a) and (b) seem like exceedingly plausible principles, and I tentatively believe that both are true; however, rejecting (PP1) does not commit us to rejecting either of these further principles. Remember what we are rejecting when we reject (PP1) is that responsibility for an action *requires* responsibility for one's reasons for acting. It can still be the case that our responsibility for our reasons in some way entails responsibility for our actions; it just can't be that responsibility for reasons is the *only* way to be responsible for one's actions.

#### IV. A Strawsonian Reply

##### *A. Identifying One Version of the Reply*

So what does Strawson have to say for himself? How might Strawson attempt to show that agent causation doesn't really solve the regress argument? Given how Strawson criticizes views similar to agent causation, I am confident that he would respond by invoking the following considerations:

Suppose  $S$  is deliberating about whether to agent cause an intention to do some action  $A_1$  for some reasons  $R_1$  or to agent cause an intention to do  $A_2$  for  $R_2$ . Suppose that  $S$  agent causes an intention to do  $A_1$  for  $R_1$ . Unless  $S$ 's choice to act on  $R_1$  rather than  $R_2$  is

*completely arbitrary* and a *non-rational flip-flop of the soul*, then *S* will have some reason to prefer  $R_1$  over  $R_2$ . Thus, even if agent causation were true, it is hard to see how rational free action is possible. (cf. 53-4 and 2002: 457)<sup>16</sup>

While I am confident that Strawson would appeal to the above considerations, I am not sure what argument he would be making. There does appear to be an argument in those considerations, but we only get vague glimpses of it. This vagueness is partially the inevitable result of adapting what Strawson does say to our present concern; however, I am inclined to believe that they inherit most of their vagueness from Strawson's original remarks, which themselves lack the precision and clarity of a formal argument. In any case, the argument being made in the above considerations isn't entirely explicit, and we have some work to do in constructing an argument from them. Perhaps the argument is supposed to be a dilemma, which we could make explicit by adding the following:

Either *S*'s choice to act on  $R_1$  rather than  $R_2$  is arbitrary [i.e. not performed for a reason] or its not. If it is arbitrary, then  $A_1$  isn't rational, and so it's irrelevant to whether humans can perform rational free actions. If it isn't arbitrary, then it is fully determined by one's reasons,  $R_3$ , which seems to imply that *S* must be responsible for  $R_3$ , if *S* is to be responsible for his choice to act on  $R_1$ .

### *B. My Rejoinder*

The above considerations, when developed into this dilemma, are clearly mistaken. I find both horns rather suspect, but let's start by rejecting the first horn. This horn will look nigh unavoidable if we fail to notice that it depends on a stronger notion of rational action than the one that Strawson himself develops and defends. What we want to know is whether  $A_1$  is a rational action, and—according to Strawson's own criteria—the answer to that question

---

<sup>16</sup> "But if it does not have any further such desires or principles of choice, then the claim that it exercises some special power of decision becomes useless in its attempt to establish freedom. For if it has no such desires or principles of choice governing what decisions it makes in the light of its initial reasons for action, then the decisions it makes are rationally speaking random; they are made by an agent-self that is, in its role as decision maker, entirely nonrational in the present vital sense of "rational." It is reasonless, lacking any principles of choice or decision" (1995, pg 27).

appears to be yes.<sup>17</sup> According to Strawson's definition of rational action, *S*'s action  $A_1$  is a *rational* action, because it is one that is fully determined by reasons  $R_1$ —and it is fully determined by *S*'s reasons even if *S* has no reason to prefer  $R_1$  over  $R_2$ . It seems the only way that Strawson could validate the first horn of the dilemma is by adopting a new and stronger notion of rational action, one asserting that an action *A* is rational only if it is *necessitated* (and not just fully determined) by *S*'s reasons *R*.<sup>18</sup> Yet why should the agent-causation theorists accept this new notion of rational action? It seems that they are *obviously* committed to denying it. For if all rational actions are *necessitated* by prior states of the agent, then rational actions can't be up to the agent—at least not in any sense that will make libertarians happy.

Maybe this stronger notion of rational action is the correct one, and if so, it would surely be a blow to agent causation theorists. Yet Strawson says nothing in favor of this stronger notion of rational action. Moreover, it seems, to me at least, that I can perform rational actions even if my reasons do not necessitate my actions; therefore, this stronger notion seems intuitively unsatisfactory. What is essential to the notion of rational action is that reasons somehow *explain* the action, and adding that the reasons must necessitate the action seems to be going a little overboard, as it were. Thus, until some rather powerful considerations are brought on behalf of this stronger notion of rational action, we can conclude that the first horn is mistaken.

The second horn is also inadequate because it seems to rely on PP1 as a premise. Since the agent causation theorist argues that PP1 is false, using PP1 as a premise would be a classic case of begging the question. Does the second horn really rely on PP1 as a premise? Appealing to

---

<sup>17</sup> I say that the answer *appears* to be yes, because Strawson only provides a partial definition of rational action, one containing only one necessary condition. So conceivably, an action could satisfy this condition, but still fail to be rational because it fails some other condition of which we are unaware. Nonetheless, I am inclined to think that something like Strawson's condition is not only necessary, but sufficient for an action's being rational. I will proceed with that assumption, and I will leave it to others to prove me wrong.

<sup>18</sup> Clarke makes some similar observations in section 9.4 of his 2003.

PP1 seems to be the only way to justify the inference from ‘some choice is fully determined by  $R_3$ ’ to ‘if  $S$  is responsible for that choice, then  $S$  is responsible for  $R_3$ .’ Hence, both horns of the dilemma appear rather lousy, and, consequently, this way of developing the above considerations is unsatisfactory. Perhaps there is some other argument which can be inspired by the above considerations, but I expect that any such attempt will not constitute an adequate response to the agent causation theorist.

### *C. Alternative Strategies for a Successful Reply*

I want to conclude this paper by mentioning two other strategies which are most likely to yield a successful reply. First, Strawson could argue that there is strong evidence that we don’t have the ability of agent causation, either because the whole notion is incoherent or there is, say, empirical evidence that we don’t have it.<sup>19</sup> If Strawson were to take this strategy, he would be implicitly conceding that his original argument doesn’t work against *all* libertarian views, as he originally claimed. He, therefore, would have to supplement his original argument with a premise asserting that we don’t have the ability to agent cause our actions.

The second strategy argues that we have more reason to believe that PP1 is true than we have to believe that agent causation would provide us with the requisite control for responsibility and freedom. This strategy might even concede that the response of the agent causation theorist provides some reason to disbelieve PP1, but it would argue that, overall, the better evidence is in favor of PP1 rather than against it.<sup>20</sup> The nice thing about this response is that Strawson wouldn’t have to supplement his original argument with new premises; he would only be providing stronger support for the original ones.

---

<sup>19</sup> This is Pereboom’s approach in chapter three of his 2001.

<sup>20</sup> Notice that this strategy would not beg the question, for PP1 would be the conclusion, not a premise, of this strategy.

In my estimation, I would say that Strawson's best chance at countering my objection to PP1 would be to adopt the first strategy. I don't find the second strategy very promising, because it is hard to see what, other than intuitive plausibility, can be adduced in favor of PP1. Yet, as I already mentioned, if Strawson takes this first strategy, he would in effect concede that his original argument does not work against agent causation, which is the very point I hope to have established in this paper.

## Bibliography

- Clarke, Randolph. 2003. *Libertarian Accounts of Free Will*. New York: Oxford University Press.
- O'Connor, Timothy. 1995a. *Agents, Causes and Events: Essays on Free Will and Indeterminism*. New York: Oxford University Press.
- \_\_\_\_\_. 1995b. "Agent Causation." In O'Connor 1995a: 173-200.
- Pereboom, Derk. 2001. *Living Without Free Will*. New York: Cambridge University Press.
- Strawson, Galen. 1986. *Freedom and Belief*. New York: Oxford University Press.
- \_\_\_\_\_. 1995. "Libertarianism, Action, and Self-Determination." In O'Connor 1995a: 13-31.
- \_\_\_\_\_. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75:5-24.
- \_\_\_\_\_. 2002. "Bounds of Freedom." In Kane, Robert (ed.), *The Oxford Handbook of Free Will*. New York: Oxford University Press.