Philosophica 85 (2012) pp. 91-121

REVISITING STRAWSONIAN ARGUMENTS FROM INESCAPABILITY

András Szigeti[*]

ABSTRACT

Peter Strawson defends the thesis that determinism is irrelevant to the justifiability of responsibility-attributions. In this paper, I want to examine various arguments advanced by Strawson in support of this thesis. These arguments all draw on the thought that the practice of responsibility is inescapable. My main focus is not so much the metaphysical details of Strawsonian compatibilism, but rather the more fundamental idea that x being inescapable may be reason for us to regard x as justified. I divide Strawsonian inescapability arguments into two basic types. According to arguments of the first type we cannot give up the practice. According to arguments of the second type we should not give up the practice. My reasons for revisiting these Strawsonian inescapability arguments are, first, to establish that these are different and to some extent conflicting arguments. Second, I hope to show that none of Strawson's inescapability arguments are convincing. Third, I discuss the possibility that the practice of responsibility is inescapable in a different, more

[*] The author wishes to acknowledge the support received for work on this article from the research project "What it is to be human?" (ERC_HU BETEGH 09) funded by the Hungarian Innovation Office (formerly NKTH)/Mag Zrt.

pessimistic sense than envisaged by Strawson. What may be inescapable under conceivable scenarios is the conflict of theoretical and practical considerations in the justification of the practice.

1. Introduction

As is well known, Strawson defends the thesis that the truth or falsity of determinism is irrelevant to the justifiability of responsibility-attributions. In this paper, I want to examine the various arguments advanced by Strawson in support of this thesis. These arguments all draw on the thought that the practice of responsibility-attributions is inescapable. I will call them "arguments from inescapability" because they move *from* the diagnosis of inescapability of the practice *to* the conclusion that the practice is justified. In fact, what I am interested in is not so much the metaphysical details of Strawsonian compatibilism, but rather the more fundamental idea that *x* being inescapable may be a sufficient reason for us to regard *x* as justified.

It is crucial to see right from the start that this is the basic underlying structure of Strawsonian inescapability arguments. Schematically, the claim is this: *T* is irrelevant to the justification of *x* for the reason that *x* is *F*. Crucially, however, this claim is made because it is thought that *x* being *F* is sufficient to justify *x*. So in our specific context the claim is that the thesis of determinism (*T*) is irrelevant to the justification of the practice of responsibility (*x*) because the practice of responsibility is inescapable (*F*). This claim is made because it is thought that inescapability of the practice of responsibility is sufficient to justify the practice of responsibility.[1] It is the plausibility of this type of

[1] It may be protested here that this misrepresents Strawson's line of thought. Strawson only says, it will be objected, that the question or challenge concerning the justification of the practice of responsibility *cannot be posed* because of the

argumentation that I would like to challenge here. However, in this paper I will restrict my critical conclusions to the debate concerning the justification of responsibility-ascriptions and leave it open whether the criticism of inescapability arguments can be extended to other areas in which such inescapability arguments are also used (e.g., anti-skeptical strategies in epistemology).

Still, it is worth noting that we find the theme of inescapability not only in his discussion of moral responsibility, but also in other areas of Strawson's work, most prominently in his proposed neo-Humean solution to skepticism (Strawson 1985).[2] Strawson's position is often referred to as "Strawsonian naturalism."[3] Strawson himself is partly responsible for this

---

inescapability of the practice (we will see in Section II below how one could make sense of the claim that the challenge of justification cannot even be posed). But I do not think this makes any difference. Note that this latter claim still entails that the practice could *not* be shown to be *not* justified by certain considerations. So the general structure of the inescapability argument would still be as before: *T* is irrelevant to the justification of *x* for the reason that *x* is *F*. So the fact that *x* is *F* is thought to be sufficient for *T* not to "unjustify" *x*. Or in our specific context: if the thesis of determinism (*T*) is irrelevant to the justification of the practice of responsibility (*x*) because the practice of responsibility is inescapable (*F*), then the fact the practice of responsibility is inescapable is thought to be sufficient for the thesis of determinism not to "unjustify" the practice of responsibility. I contend that this way of arguing is as controversial as the one discussed above. In any case, the objections made later on apply to it as well.

[2] It is worth noting that the words "inescapable" and "inescapability" or synonymous expressions occur with remarkable frequency in the work referred to above. The connection between Hume's thought and Strawson is explored among others in Williams 1996, see xiii, 11-5, 24-5, etc.

[3] For example, Sher talks of Strawson's "uncompromising naturalism" and describes his theory as "relentlessly naturalistic". See Sher 2005, 81, 85.

somewhat misleading label as he refers to his preferred way of meeting the skeptical challenge in epistemology, morality and elsewhere as "non-reductive naturalism" (Strawson 1985, 24, 39-41). However, as will be seen, the argument from naturalism is just one way for Strawson to highlight the importance of inescapability and, specifically, to highlight the supposed connection between inescapability and justification. I will try to show in the following that Strawson in fact advances a number of separate and to some extent conflicting arguments to support the idea that inescapability can justify.

In my reading, Strawson himself lays out at least four inescapability arguments with regard to the practice of responsibility (as some of these arguments can be interpreted to mean quite different things, we will find that there may well be even more than four inescapability arguments). One of my reasons for revisiting these Strawsonian inescapability arguments is to show that although they all appeal to the idea of inescapability these really are quite *different* arguments.[4] In this respect this paper serves as an exegetic exercise. In addition to this, I also hope to show, second, that in their original formulation all of Strawson's inescapability arguments are unconvincing. And finally, third, I want to discuss the possibility that the practice of responsibility may be inescapable in a different, more pessimistic sense than envisaged by Strawson.

I propose to divide the Strawsonian inescapability arguments into two basic types. Arguments 1 and 2 below belong to the first type. According to these, determinism is irrelevant to the justifiability of responsibility-attributions because we *cannot* give up the practice. Arguments 3 and 4 below belong to the second type. These arguments appeal not to the incapacity to give up the practice, but to practical rationality. The idea is

[4] Perhaps exactly because they recognize the potential conflict between different Strawsonian arguments, many authors commenting on Strawson focus only on a subset of these arguments at the expense of others.

that whether or not we could give up the practice of responsibility we have overwhelmingly strong practical reasons why we *should not* do so.

In the next four sections (Sections II-V), I will reconstruct these four arguments and spell out my main objections to each of them. Finally, in Section VI, I will propose yet another sense in which the practice of responsibility may be taken to be inescapable. I will build here on what I will refer to as the "Wiggins-conjecture". As will be seen, this understanding of inescapability diverges from the original Strawsonian arguments – most importantly in that it no longer subscribes to the idea that the inescapability diagnosed could be used as a form of justification. Nevertheless, I hope to show that it may still be the philosophically most fruitful reading of the Strawsonian notion of inescapability, a reading which, in addition, is not totally alien to the spirit of Strawson's groundbreaking thoughts on this topic.

2. Argument one: No justification

According to the first argument (Argument 1), the demand for justifying the practice of responsibility as a whole is misguided. Such a wholesale justification is neither possible nor necessary: "the existence of the general framework itself neither calls for nor permits an external reaction justification" (Strawson 1985, 41; see also Strawson 1974, 23[5] and Hieronymi 2004). If that is true, however, it is simply irrelevant whether or not the truth of determinism would *per impossibile* require us to give up the practice of responsibility-attributions.

What does this suggestion mean exactly? Already in the quote taken from Strawson we find two quite different ideas as regards the justification of the general framework of the practice: first, that the

[5] The wording of the relevant passage in this text is almost but not completely identical to that quoted above.

general framework does not "call for" justification, and second, that it does not "permit" justification. Justification not being called for is not the same as justification not being permissible. Moreover, each of these ideas themselves may be interpreted in different ways. Let me therefore take these ideas one by one to help us better assess the first argument from inescapability.

2.1 No justification necessary

So the first idea is that the general framework does not "call for" justification. Why this may be so could be explained in a number of ways. There is some textual evidence in Strawson's pertaining works for each of the possible explanations I will now discuss.

Thus, first, one could explain why the general framework does not call for justification by arguing that it is *non-rationally* grounded. The point here would be that the practice of responsibility is, to quote Hume: "more properly an act of the sensitive, than of the cogitative part of our natures" (Hume 1978, 183). If so, then the demand to justify the practice of responsibility-attributions may be like asking us to justify our most basic emotional propensities and reflexes. Such justification is not called for because it is pointless.

To fully appreciate this interpretation of the "no justification argument" we need to recall Strawson's preoccupation with the role emotions play in the practice of responsibility. As is of course well known, one important novelty of the Strawsonian approach is the emphasis on the connection between the concept of responsibility and the world of emotions. Strawson has shown that this connection is not just a contingent feature of human psychology. But if it is true that certain human emotions constitute the general framework of the practice of responsibility and if it is also true that at some basic level these emotions are hard-wired, then we would have an explanation why the general framework itself does not call for justification.

Even if Strawson is right, however, that this connection between the practice of responsibility and the world of emotions is not contingent, it remains a moot question how decisive the connection really is. First, it is not clear that attributions of responsibility are necessarily bound up with affective states. But even if they are, second, emotions are responsive to reasons. We can be reasoned out of our affective states. Strawson is clearly aware of these considerations of course.[6] For example, at one point he talks about the "whole web or structure of human personal and moral attitudes, feelings, and *judgments*" (Strawson 1985, 39 – my italics). For these reasons, a purely emotivist interpretation[7] along these lines is not only insufficient to bolster the "no justification argument", it would also constitute a coherent but overly simplistic reading of what Strawson himself says.

What other explanation may there be of why the general framework itself does not call for justification? Here is a different suggestion. This is again based on things Strawson himself says, but it does not require a non-cognitivist framework. The idea here is that the general framework does not call for justification because there is simply no reason for us to question the validity of that framework.[8]

---

[6] This is the reason why, in my opinion, Galen Strawson's characterization of his father's account as a "non-rational commitment theory of freedom" (see Strawson 1986, 84 and *passim*) is misleading.

[7] Such as that offered by Jonathan Bennett in an otherwise wonderfully insightful essay on Strawson. For example, Bennett says that Strawsonian reactive attitudes express "my emotional make-up, rather than reflecting my ability to recognize a blame-meriting person when I see one" (Bennett 1980, 24) and in the same vein attributes (fallaciously in my opinion) the view to Strawson that "reactive feelings cannot be made impermissible *by any facts*, e.g., the fact that men are natural objects" (Bennett 1980, 29 – my italics).

[8] Several commentators emphasize in their reading this alternative strategy at the expense of other arguments which as I am trying to show are also to be found in Strawson. See, for example, Stern 1974, 73: "The question whether it is rational

It is true as Strawson readily concedes that we can regard human actions from a perspective that lies outside the general framework of the practice of responsibility. This is the perspective of the objective attitude. From this perspective we characterize human behavior in purely naturalistic-causal terms "which exclude moral praise or blame" (Strawson 1985, 50). At the same time, for Strawson the existence of this other framework does not call into question the validity of the general framework of the practice of responsibility. These alternative standpoints are not incompatible because they do not conflict. Neither of them is more correct or more real than the other. Each is real and correct relative to its own standards (Strawson 1985, 45).[9] And so justification of the general framework of the practice of responsibility is not called for because there are no standards based on which the validity of this framework could be challenged. All meaningful questions regarding justification arise first *within* this framework (Strawson 1974, 23).

I believe that this alternative strategy to save the "no justification argument" will not work either. It is not true that two standpoints do not conflict. The objection I am making here is essentially the same as that made against the similar (but more comprehensive) Kantian idea of "insulating" the practical from the theoretical perspective.[10] The anti-Kantian point was that the question of causality appears to be directly pertinent to the practical perspective. By the same token, what seems

---

to give it [the commitment to reactive attitudes] up cannot even be raised: rational justification takes place within the framework of basic human commitments."

[9] Strawson draws an analogy here with human perception (Strawson 1985, 45-46). Seen from the ordinary human standpoint blood is red, viewed under a microscope it is colourless. But there is no reason why we should claim one perspective to be more real or more "justified" than the other.

[10] On the Kantian "insulation strategy", see Wallace 2006, 159-64.

wrong with the Strawsonian version of the insulation strategy is that how we explain an action in causal terms will be very much relevant to whether ascribing responsibility for that action is justifiable or not.

We see this once we ask why responsibility-undermining conditions should be pertinent to the justifiability of ascriptions of responsibility. The analysis of excuses such as coercion or mental deficiency shows that whether the agent could have done otherwise is not merely a consideration relevant to the theoretical perspective.[11] On the contrary, it is very much relevant to whether it is *morally* appropriate to ascribe responsibility to the agent for that action. If the agent could not help doing what she did, the action will not only be mistakenly described, but the agent herself will be wronged.

If this is correct, then it has not been shown that it is not possible to call into question the validity of the general framework of the practice of responsibility-attributions from an alternative, external perspective. If that external perspective reveals that no one could ever help doing what they did, then it will always be morally wrong to ascribe responsibility for any given action. We should also note here that it would be surprising if the situation were otherwise. After all, calling the general framework of responsibility into question is precisely what, among others, hard incompatibilists do![12] We may disagree with them, but we do not feel that their perspective on responsibility is conceptually confused.

I conclude that this strategy for defending the "no justification argument" fails as well. Emphatically, what this critical conclusion entails is *not* that the general framework of the practice of responsibility could not in principle be compatible with another external framework,

[11] *Pace* Korsgaard 1996, esp. 197-8.

[12] Furthermore, it is possible to mount such a wholesale challenge to the practice of responsibility on *moral* grounds as well. Thus some complain that our practice of responsibility is vindictive and excessively punitive. See, for example, Baier 1995 and Wertheimer 1998.

even the framework of a world of deterministic causation. What is rejected here is only the Strawsonian idea that the general framework of the practice of responsibility is *necessarily* or by definition immune to challenges from a perspective outside this framework.[13]

This objection also takes care of the related Strawsonian strategy which consists in arguing that the general framework does not call for justification because this framework serves as the "scaffolding", "background", "substratum" (these are originally Wittgenstein's metaphors, but Strawson 1985, 20, 28 quotes them with approval). What is suggested here by Strawson is that there is no platform, no perspective from which to carry out the justification of the framework itself. To use Gary Watson's phrase, there is simply "no more basic belief" (1987, 255) to appeal to in order to justify the general framework of the practice of responsibility. If the above objection is correct, however, then there can be such more basic beliefs, namely those concerning determinism (as well as those concerning other necessary metaphysical conditions). The truth of these basic beliefs may indeed call into question the entire general framework of the practice of responsibility.

[13] There is one passage where Strawson tentatively introduces something like an "argument from illusion" (Strawson 1985, 50). His point is that if the external perspective did undermine the practice of responsibility, then it would follow that "we live most of our lives in a state of unavoidable illusion" unless we assumed that the two perspectives did not conflict. Therefore, we should assume that the two perspectives do not conflict. Of course, since then Saul Smilansky (see Smilansky 2000) defended the view that it is precisely such a state of illusion in which we live most of our lives. Once again, Smilansky may be wrong, but it does not seem like his position is conceptually impossible.

2.2 No justification possible

It will be remembered that the other Strawsonian idea I distinguished above as part of the "no justification argument" was that the general framework does not "permit" justification. Again, why this may be so could be explained in a number of ways. Again, there is some textual evidence in Strawson for each of the possible explanations I will now discuss (with the possible exception of the last idea to be mentioned in this section). And again, these explanations are not synonymous despite the fact that Strawson does not really keep them apart.

Thus the general framework would not "permit" justification if, for example, the general demand for justifying the practice of responsibility-attributions was somehow *self-refuting.* This is, I think, what Strawson had in mind when proposing the following well-known argument (Strawson 1974, 11): Ascriptions of responsibility are undermined when something abnormal is true about the action or the agent. If the whole practice of responsibility (i.e., the general framework) were unjustified, then all actions and agents would be abnormal. But as a matter of logic, abnormality cannot be a universal condition.

As Paul Russell has shown, this version of the argument equivocates between abnormality and incapacity (Russell 1992). While abnormality cannot be a universal condition, incapacity can be. If determinism is true, it may well be that we are all and always in the relevant sense incapacitated. To show that we are not may well be possible, but it requires further argument. So not only does it seem wrong to say that the general framework does not permit justification, the general framework positively requires such justification!

Finally, I would like to broach an idea for which I have not found explicit textual evidence in Strawson. I suspect, however, that this idea may also lurk in the back of the minds of those (and perhaps Strawson is one of these people) who think that the "no justification argument" must be right because asking for such a wholesale justification would be self-

refuting. According to this last idea, the demand could be self-refuting in terms of the kind of ethical concerns we can have.[14] The suggestion is that we cannot be morally worried about the demand without being already committed to the practice. For what would be the case, if the practice were *not* justified? We would do wrong by ascribing responsibility to one another for our actions. But it would not matter that we did wrong in this way *unless* we were committed to hold ourselves responsible for doing wrong. In short, once we raise the demand for justifying the practice of responsibility-attributions, we are already part of that practice.

I think we should reject this suggestion too. First, the demand for justifying the practice is not driven by an ethical interest alone. We can have theoretical reasons for insisting on this demand too. Second, we can be worried for moral reasons about the wrong involved in attributing responsibility even if the practice of responsibility-attributions turns out to be unjustified. If the practice of responsibility-attributions is unjustified, we do each other harm by blaming, punishing, etc. ourselves

[14] We find this idea for example in Gosepath 2009, 267: "Anders als im Fall blosser Autorität, wo wir stets hinterfragen können, warum wir handeln sollen, ist diese Rückfrage mit Bezug auf das Bewertungsschema als solches sinnlos. Selbst das Hinterfragen des Schemas ist Teil dieser Praxis, die durch das Schema strukturiert wird." ["Unlike in the case of pure authority which we can always challenge as to why we should act, such a challenge of the evaluative framework is pointless. This is because that very challenge of the evaluative framework forms part of the practice which is structured by the framework."] Of course, no reference is made here to Strawson. Further, the passage concerns the totality of the practical domain of normative reasons. But the underlying thought is the same. The justificatory demand makes sense only within the general framework, not outside it: I must leave the question open here whether this thought is more plausible for the *entire* practical domain (as is suggested in the passage quoted from Gosepath) than for a mere subset of this domain such as the practice of responsibility (which is what the Strawsonian argument focuses on).

for our actions. Even if no one is to be held responsible for this, we can find all this morally repugnant.

3. Argument two: Naturalism

I now come to the argument from naturalism (Argument 2). The crucial premise of this argument is that the commitment to responsibility-attributions is a natural fact. It is a deeply ingrained part, a "given" of human nature (Strawson 1974, 18, 23 and Strawson 1985, 33, 39). Attributing responsibility to others and ourselves, praising and blaming are "natural expressions of natural responses to what we see people do" (Wolf 1981, 389). If indeed there is a thoroughgoing psychological incapacity rooted in human nature which makes it impossible for us to give up the practice of responsibility-attributions, then it is quite irrelevant that the thesis of determinism would have such an implication.

How does this naturalistic argument differ from the "no justification argument"? First, the argument from naturalism is not intended to demonstrate the non-rational character of our commitment to the practice of responsibility. It is true that the plausibility of this argument too becomes more forceful when coupled with Strawson's other main point about emotions playing an important and non-contingent role in the practice of responsibility. Having said that, Strawsonian naturalism need not entail assumptions about the non-rational or emotivist basis of the practice of responsibility. It is worth repeating that what we are by nature committed to according to Strawson is "whole web or structure of human personal and moral attitudes, feelings, *and judgments*" (Strawson 1985, 39 – my italics).

Second, the naturalistic argument differs from those other cognitivist versions of the "no justification argument" discussed above as well. These latter versions of the argument were all based on the idea that the demand for justifying the practice of responsibility as a whole involves

some sort of conceptual or even logical confusion, a misunderstanding as regards the scope of the justificatory demand (see Section 2 above). The argument from naturalism by contrast is based on a point about the *fundamentum in re* of the practice.

I think that the point of the argument from naturalism can be best understood by asking why a natural fact is thought to justify the practice of responsibility. As I said, Strawson's answer is that what invoking this natural fact gets us is the recognition that it is impossible for us to give up the practice of responsibility. But if it is impossible to give up the practice, then it is in vain to argue that we should. There is an "ought" only where there is a "can". We have not chosen our commitment at the first place, nor can we choose to opt out of it. Hence arguments purporting to produce reasons why we should do so are as idle as arguments as to why we should aim to have eternal life. Inescapability justifies.

The first important objection to this line of thought is the following. It may well be true that we have a general natural inclination to attribute responsibility to one another. At the same time, we are able to check this natural inclination in any given token case (see Russell 1992). So even if the naturalist is right about what *types* of reactions characterize human nature, this has no bearing on how we are able to respond in any given *token* situation. If this is correct, then it has not been shown that attributions of responsibility are inescapable in the required sense and hence that determinism would be irrelevant. This is because now we see that we *can* respond to the recognition that the action was determined in any given token case by not attributing responsibility for that action.

In other words, the *descriptive* premise of Strawson's argument from naturalism appears to be simply false. It is not true that in specific situations we would be inescapably committed to attributing responsibility as a sort of naturalistic reflex. Our naturalistic commitment is at best a tendency or inclination. As such, it has no bearing on what it is justified to do or to believe in particular situations.

But we can go even further and formulate a second, *normative* objection as well. For the sake of the argument, consider the possibility that Strawson's descriptive premise was correct. If so, then there would indeed be token cases in which we could not refrain from ascribing responsibility. Imputing responsibility would be like a reflex or a gut reaction that we could not always control. Would in such cases the mere inescapability of the response justify the ascription of responsibility?

I think not. In fact, I think Strawson would be forced to admit this himself based on what he says elsewhere about responsibility-undermining conditions. He makes it quite clear in those passages that whenever an excuse or exemption obtains – i.e., when the agent could not help doing what she did for some reason – we are as good as morally *required* to check our ordinary reactions and withhold our attributions of responsibility (see Strawson 1974, 7-9). But if that is true, then inescapability of the reaction cannot be sufficient to justify it. For a responsibility-undermining condition may well obtain and our reaction of attributing responsibility may not be justified despite the fact that the reaction was inescapable.

I conclude that the argument from naturalism is based on a descriptively implausible hypothesis. Moreover, even if it were not descriptively implausible, we find that inescapability as a fact of nature is not a consideration which in itself would be sufficient to justify either a practice as a whole or particular instances of applying this practice.

4. Argument three: Value

The discussion so far has shown that it is neither logically confused nor psychologically impossible to hold that we *can* escape the practice of responsibility. So perhaps we need to look for a different sense in which this practice may be inescapable. Perhaps a more feasible idea to consider is that even if strictly speaking we could escape the practice of

responsibility, there is no reason why we *should.* The practice would now be understood to be inescapable in the sense that given our values and practical commitments – that is, "in the light of an assessment of the gains and losses to human life, its enrichment or impoverishment" (Strawson 1974, 13) – we cannot conceive of a good enough reason to escape it.

Strawson formulates this thought in two different arguments. These will be discussed in this and the following section. The first of these is the *argument from value* (Argument 3). This holds that we do not have a reason to abandon the practice of responsibility because this would amount to an inacceptable loss of things we value, an unbearable impoverishment of human life.

Again, there are two ways to make sense of how abandoning the practice of responsibility could be thought to lead to such an impoverishment of human lives. On the stronger version – let us call it the "all-in version" – *everything* that is of value, and perhaps valuing itself, depends on the practice of responsibility. The all-in version is defended for example by Susan Wolf, an author with strong Strawsonian sympathies: "living in accordance with the fact that we are not free and responsible beings would require us to give up *all our values"* (Wolf 1981, 401-2 – my italics).

Surely, this is an exaggeration. The practice of responsibility cannot be a necessary condition for the existence of all values. It is hard to see why all valuable things would cease to be valuable in a world without responsibility. Would works of art cease to be beautiful? It is no less hard to see why there would cease to be *moral* value or disvalue in such a world. Would it not still be horrendous to set a cat on fire?

Also, many argue that some cultures lack the notion of responsibility at issue here. Wolf's claim would commit us to saying that such cultures cannot exist. For the idea of a culture without at least some values is obviously incoherent, but according to Wolf without responsibility there

can be no values and so there cannot be cultures without our notion of responsibility.

I conclude that it is quite clearly false to say that values in general presuppose the attribution of responsibility. Even the more restricted claim, that our valuing of human actions or activities presupposes the attributability of responsibility, is doubtful. Even if Mozart's creative output was entirely determined by his genes, I can still admire his works and the musical genius producing them.

But perhaps *some* values do presuppose the practice of responsibility. If so, then the argument from value would be based on the claim that abandoning the practice of responsibility would impoverish human lives because it threatens the loss of this specific subset of our values. And indeed, there is a good case to be made that without treating ourselves as responsible agents we would have to abandon, or revise beyond recognition, guilt, forgiveness, punishment, praise, admiration as well as some norms of distributive fairness.

There is no place here for a detailed discussion of how the practice of responsibility can bring forth or constitute distinct forms of value. The following examples should suffice, however, to prove that there is such a connection between the practice and certain specific values.

So, first, it seems quite plausible to regard forgiving as something valuable. It is also quite plausible to say that forgiveness presupposes responsibility in the sense that you can only forgive me for what I have done to you if you think that I was responsible for my action at the first place (Kolnai 1974). But then it follows that realizing the value of forgiveness presupposes that the practice of responsibility is in place.

Second, it is also quite plausible to regard indignation as something valuable (whereby indignation is defined here as moral reaction to harm caused by an agent to a third party to whom one is not personally related in any way). Surely, it is valuable that we can be upset about something Person *A* does to Person *B* even if this action has absolutely no harmful consequences for us. Again, it is quite plausible to say that such

indignation presupposes the attribution of responsibility.[15] And so again, it follows that realizing the value of such kind of sympathetic indignation presupposes that the practice of responsibility is in place.

If so much depends on the practice of responsibility, we may indeed be significantly worse off without this practice. This would give us an overwhelmingly strong reason not to give up the practice of responsibility-attributions.

Now, as is well known, many disagree with the claim that the practice of responsibility can be a source of value. With or without a hard incompatibilist background,[16] many argue that we would be better off without the practice. However, it is crucial to see that we need not embrace such a negative assessment of the practice of responsibility in order to reject the argument from value.

In my view, it is one of the merits of the Strawsonian theory of responsibility to have called attention to the fact that we do well by the practice of responsibility. This practice can indeed be a source of value in human lives in ways described in the two examples just given. The question is only what follows from this as regards the justifiability of the practice. For once the practice of responsibility is just one source of value among others (or more precisely, the source of a specific subset of values), and not a precondition of the existence of every value, as Wolf would have us believe, it no longer follows that the practice is inescapable. The price of abandoning the practice may still be significant, but it is no longer *prohibitively* high.

---

[15] Because indignation (or at least some distinctly recognizable kind of indignation) is triggered by the thought that Person *A* caused harm voluntarily to Person *B* and can therefore be held responsible for her action.

[16] For the former, see Pereboom 2007, for the latter, see Baier 1995 and Wertheimer 1998.

There can be all kinds of reasons why we would be willing to pay such a price.[17] We may agree, for example, that forgiveness is something valuable, but come to the conclusion that the loss of this value is offset by the gain of overcoming our culture's obsession with guilt – guilt being another reaction that presupposes the attribution of responsibility.[18] We may agree too that indignation is something valuable, but decide that the loss of this value is offset by the gain of embracing a more enlightened regime of sanctions that does not presuppose the attribution of responsibility unlike some traditional forms of punishment. And so on.

---

[17] I cannot go into two particularly interesting questions here, namely who this "we" really is, and how, historically speaking, comprehensive revisions of an existing moral practice can occur. But let me just briefly gestures towards some of the issues at stake. As regards the first question, we may wonder for example how central the practice of responsibility really is to modernity? Is the "we" of modernity co-extensive with the "we" of participants in the responsibility-practice? Or can different cultural or religious traditions co-existing in our age, or say even different countries, institute the practice in significantly divergent ways? As regards the second question, we need to ask what explains changes which a moral practice, or an important part thereof, can undergo throughout history? Further, are there some parts of moral practice which are (necessarily?) universal? Obviously, I cannot go into these questions here, but see Williams 1976 and Williams 1993. See also my discussion of Susan Wolf's position earlier on.

[18] Perhaps not all forms of guilt presuppose the (self-)ascription of responsibility. And perhaps not all forms of guilt presuppose the (self-)ascription of responsibility for a *voluntary* action. I cannot argue this point here, but I am in agreement with those authors (e.g., Rawls 1971, 482) who think that there is a distinct form of guilt which is experienced because and only because one assumes responsibility for something one has done voluntarily. I believe that, if anything, it is this form of guilt our culture may be accused of being obsessed with (for a discussion of and qualified support for this charge, see Williams 1993).

And finally, once the practice of responsibility is not inescapable, even purely theoretical considerations can seem to constitute a good enough reason to abandon the practice. We may just want to "live in accordance with the facts" (see Wolf 1981, 393). If these facts on our best theory of the world turn out to be such that the practice of responsibility no longer appears to be justifiable, this too can be a good enough reason to escape the practice.

It also follows from this too that the argument from value cannot be used to write off determinism as irrelevant to the justifiability of the practice. If the truth of determinism is found to undermine the justifiability of the practice of responsibility, then the truth of determinism too may seem to constitute a good enough reason to opt out of the practice even if this practice is something we rightly value.

5. Argument four: Rationality

At the beginning of the previous section, I said that we find two different arguments in Strawson in support of the idea that we cannot conceive of a good enough reason to escape the practice of responsibility. The gist of the argument discussed in the previous section was that we could not possibly think of a *strong enough* reason to abandon the practice of responsibility because of the loss of value this move would involve.

However, perhaps Strawson's idea is rather that certain considerations, such as the truth of determinism, are not the right *kind* of reason to abandon the practice of responsibility. This idea forms the basis of the *argument from rationality* (Argument 4). The point here is that a metaphysical, that is, a theoretical description of the world, however

accurate, cannot generate practical[19] reasons for us to abandon the *practice* of responsibility-attributions.

Determinism is such a "general theoretical doctrine" (Strawson 1974, 13). Therefore, it can only have theoretical implications (whatever these we may speculate to be), but not practical consequences. As a theoretical thesis, determinism must remain irrelevant to our *practical* choices. In short, the practice of responsibility remains inescapable *in practice*, even if the metaphysician should seek to escape it by means of (practically) idle theoretical speculations. Or so it is argued.

This would also explain the somewhat cryptic statement made by Strawson in a footnote, namely that even if we could choose to do so "it would not necessarily be rational to choose to be more purely rational than we are" (Strawson 1974, 13n1). One way to dissolve the appearance of paradox is precisely to use the argument from rationality discussed here. The choice we would be making to abandon the practice of responsibility ("if such a choice were possible") would be a practical one. Provided the argument from rationality is correct it would not diminish the practical rationality of this choice to ignore a purely theoretical consideration in making it.[20]

However, I think the argument from rationality must be rejected as well. I offer two objections against this argument below.

---

[19] It is hopefully already clear from the discussion so far, but it is worth emphasizing once again: By referring to "practical reasons" and "practical justification", I do not mean (and of course Strawson does not mean either) merely pragmatic considerations, as a "white lie" may be justified by pragmatic considerations. Rather, practical justification is to be based on our basic moral (and broader normative) concerns.

[20] This also throws light on Strawson's claim that both optimists and pessimists about the justifiability of responsibility tend to "overintellectualize the facts". They overintellectualize by paying too much attention to irrelevant theoretical considerations such as the truth or falsity of determinism.

First, as already mentioned, the Strawsonian account recognizes of course that standard responsibility-undermining conditions (e.g., coercion, ignorance) are relevant to the justifiability of ascriptions of responsibility. Now we have said before that a plausible way to account for these responsibility-undermining conditions is to say, first, that they undermine responsibility because they indicate that the agent could not help doing what she did. And then it is also plausible to add, second, that it is *morally wrong* to attribute responsibility when an agent could not help doing what she did.

However, and again this a recurring point as well, the thesis of determinism can also be quite plausibly read as implying that the agent could not help doing what she did. So if it is true that it is morally wrong to attribute responsibility to the agent when an agent could not help doing what she did, then it is morally wrong to attribute responsibility to the agent if determinism is true.[21] Therefore, determinism could have straightforwardly practical implications whether or not it is a "general theoretical doctrine". If so, then it is quite false to say that practice of responsibility is inescapable in the sense of being immune from certain theoretical challenges. Theoretical considerations may well constitute just

[21] This line of argument can be challenged in a number of ways of course. First, as is mentioned in the next paragraph above, it can be argued that determinism does not imply that the agent could not help doing what she did in a responsibility-undermining sense. But, more interestingly, second, it can also be argued that we accept responsibility-undermining conditions not for the reason that they would imply that the agent could not help doing what she did. Wallace (1994), for example, argues that standard responsibility-undermining conditions are, contrary to the standard view summarized above, *not* based on the consideration that the agent could not help doing what she did, but rather on quite different moral principles of fairness. If Wallace is right, then even if the truth of determinism did entail that the agent could not have done otherwise it may well be morally acceptable to ascribe responsibility to the agent even if determinism was true.

the kind of reasons that would call into question our adherence to this practice.

It may be argued that determinism is different from standard responsibility-undermining conditions. Needless to say, many compatibilists think precisely that holding that determinism does not entail that the agent could not help doing what she did, not at least in a responsibility-undermining sense. They may well be right. Whatever the merit of their arguments, however, these arguments will not be based on the idea that determinism is a theoretical thesis and as such cannot have practical consequences. Quite the contrary! Compatibilists will typically defend their position on the basis of theoretical considerations concerning what the thesis of determinism implies and what it does not imply.

My second objection is that the Strawsonian argument from rationality unacceptably "loads the dice" in favor of practical justification. Let us assume for the sake of the argument that Strawson is right: the practice of responsibility is inescapable. Surely, there is no denying that this bare fact would have theoretical consequences! No matter what shape the practice of responsibility will exactly take the very existence of this practice will bear on the sort of freedom we have, the constitution of our agency, and so on.

So the upshot of this second objection is that if practical facts can have such metaphysical implications, it seems awkward to deny that metaphysical considerations, such as the thesis of determinism, can have practical implications too. But if determinism can have such practical implications, then once again we find that the practice is not inescapable in the sense of being insulated from the impact of theoretical considerations. As noted, these may well be such that they would justify opting out of the practice of responsibility.

6. Inescapability without justification

It follows that we should reject the original Strawsonian program of excluding from the justificatory project certain metaphysical considerations as irrelevant non-starters. If the challenge of justifying the practice of responsibility is to be met, both normative and metaphysical concerns must be addressed.

At the same time, there may well be a grain of truth in the Strawsonian argument from rationality, the last type of inescapability argument discussed above. This grain of truth is that metaphysical and normative concerns about the practice of responsibility arise to a large extent independently from one another. Normative concerns arise when we begin to look for a notion of responsibility that is morally defensible, or better, one which forms part of a defensible moral outlook. The metaphysical concern presents itself when we investigate how agency forms part of the fabric of the world.

The problem is that these concerns, though arising independently from one another, mutually bear on each other. The main upshot of my criticisms of Strawsonian inescapability arguments has been precisely that the theoretical and practical perspectives cannot be insulated from one another. Theoretical considerations will necessarily have normative implications. And conversely too, what appears to us to be practically justified will necessarily bear on our theoretical commitments. At the same time, these perspectives differ in their aims: from the theoretical perspective we seek to construct the most convincing metaphysical picture of the world whatever the normative implications of this picture may be. By contrast, as just noted, the aim of the practical perspective is to find moral (and more broadly normative) justification for the practice of responsibility.

But if the aims of the metaphysical and ethical "projects" regarding the practice of responsibility continue to differ, then it must at least be *possible* that the two perspectives will conflict. That is to say, it should not be taken for granted that the metaphysically speaking best supported description of the world allows for the morally speaking most defensible practices. In the remaining part of this essay, I want to investigate this possibility and its relation to Strawsonian ideas.

It was David Wiggins who first broached the idea that the possibility of such a conflict could be a potential implication of the Strawsonian theory. Wiggins (2002, 300[22]) says the following: "Maybe, even if the falsehood of the [non-deterministic] assumption were authoritatively revealed (i.e., if it were authoritatively revealed that strict determinism […] obtained), it would *still* be rational for us to maintain the practices that are conditioned by the assumption. Strawson would have managed to show that there were overwhelmingly good rational reasons – reasons that even outweigh the concern for truth – for us to distract our own attention from the falsehood of the non-deterministic assumptions that conditions our practices." Wiggins, however, never really unpacks this intriguing idea. How can we have "good rational reasons that even outweigh the concern for truth"? Does this mean that a practice could be justified[23] even though it is metaphysically impossible? And conversely – we can now add to Wiggins's original surmise – could the metaphysically acceptable theory of agency yield morally unjustifiable conclusions? Let us call this, i.e., the conflict of metaphysical and practical justifiability, the Wiggins-conjecture.

---

[22] Note that Ayer makes a similar suggestion in Ayer 1980, 12-13.

[23] Whereby, to repeat a point made earlier (see note 19 earlier), practical justification is not a pragmatic, instrumentalist one of the kind we would employ to justify a white lie. Rather, such practical justification is anchored in our fundamental moral convictions.

I want to argue that the Wiggins-conjecture signals a real possibility. For example, we may well find that the morally justified practice of responsibility is predicated on a desert-based, retrospective notion of responsibility rather than a forward-looking consequentialist notion (this, of course, is also Strawson's view). There is no guarantee, however, that the required metaphysical underpinnings of this desert-based practice will be theoretically defensible. And conversely too, we may well find that the compatibilist picture of agency presupposed by the forward-looking consequentialist notion of responsibility is the best metaphysical theory we can come up with. But this is no guarantee that the consequentialist notion of responsibility will be morally justifiable.

There are various authors, especially in the recent literature, who explicitly deny that such a conflict is possible *in principle.* For these authors, the Wiggins-conjecture is necessarily false. This approach differs from traditional theories of responsibility which often simply assume that ethical and metaphysical considerations will just happen to converge somehow at the end of the day. I will discuss "neo-Strawsonianism" as an example of the new approach below. But first consider the traditional methodology.

Classical consequentialism is a good example of the old-school method. This view steers clear of the conflict by insisting, first, that the consequentialist forward-looking concept of responsibility, according to which ascriptions of responsibility serve to deter and encourage, is morally speaking the most attractive option because it avoids the morally despicable idea of "natural retaliation for past wrong [which] ought no longer to be defended in cultivated society" (Schlick 1962, 152). And more generally, we may add, it is morally the most attractive option because it is in accordance with the consequentialist principle of maximizing expected utility. And second, consequentialists also point out that, as it happens, this forward-looking notion is from the metaphysical perspective the only viable concept because it is the only one compatible

with the explanation of human behaviour in terms of causal laws of the natural world (Schlick 1962, 144).

So vintage consequentialism would be an example of a two-tiered theory of responsibility in which the metaphysical and ethical tiers happen to fit together as if by a fortunate coincidence. But the idea of those who think that a conflict between the ethics and metaphysics of responsibility is in principle impossible is different. The idea here is that one or the other domain *necessarily* enjoys priority and so a conflict between ethical and metaphysical considerations is a conceptual impossibility.

For example, a number of neo-Strawsonians (see for example Dennett 1984, Wallace 1994, Vargas 2004) defend the view that the ethical domain enjoys priority in justifying the practice of responsibility.[24] These authors diverge from Strawson in that they abandon the idea that the practice of responsibility would be inescapable (for example, Wallace explicitly denies this, see Wallace 1994, 31-32 and elsewhere). At the same time, they retain the idea common to the Strawsonian "argument from value" and the "argument from rationality" discussed above, namely the idea that we have to turn to normative features of the practice of responsibility to justify it. The practice is justified, it is said, because it is supported by general moral considerations such as requirements of

[24] But one does not have to be a Strawsonian to argue in this fashion. For instance, van Inwagen insists that "it is […] evident that moral responsibility does exist: if there were no such thing as moral responsibility nothing would be anyone's fault, and it is evident that there are states of affairs to which one can point and say, correctly, to certain people: 'That is *your* fault.' " (van Inwagen 2008, 328). If the existence of such a thing as moral responsibility is incontrovertible, however, then it follows that so long as libertarianism is false, compatibilism *must* be right. This conclusion is in fact accepted not only by van Inwagen but by several libertarians as well who say that if they were convinced of the truth of determinism they would immediately become compatibilists (rather than accept a position such as Pereboom's hard incompatibilism).

fairness (Wallace). Or the practice is justified because it represents what is "worth wanting" (Dennett).

But the question is why moral considerations should be prioritized in this way. How could moral considerations settle metaphysical disputes? Wallace (1994, 85), for example, implicitly answers this question by criticizing "metaphysical interpretations [which] postulate facts about responsibility that are completely prior to and independent of our practice of holding people responsible". He thinks that: "these interpretations seem unpromising, since it is hard to make sense of the idea of a prior and independent realm of facts about moral responsibility". I would object that it does not seem hard at all to make sense of a realm of such facts. They are of course not independent in the sense that they concern the necessary preconditions for justified attributions of responsibility. But they are independent in the sense, *pace* Wallace, that they are not constituted or brought into existence by the practice of responsibility.

Similar questions could be asked about attempted justifications on the basis of what is "worth wanting". The point here is not just that we may not agree with Dennett about what is worth wanting. Rather, the point is that we should not exclude the possibility that perhaps what is worth wanting is not how things (metaphysically) happen to be.

The converse is also true, however. Metaphysical theories certainly bear on moral disputes, but they are not somehow endowed with a special authority to settle these disputes. It is no coincidence, for example, that after defending hard incompatibilism on various metaphysical grounds against its libertarian and compatibilist rivals Pereboom goes on to present moral arguments to prove the ethical advantages of his hard incompatibilist position (Pereboom 2007, 114-124). This move is prompted by the recognition that it is by no means obvious that the metaphysically most convincing picture of freedom and agency will satisfy our ethical intuitions concerning responsibility. Pereboom, of course, thinks that ultimately they will. What I have argued above is that we should also consider the possibility that they will not. In other words,

neither metaphysical nor ethical considerations can trump each other just by virtue of being metaphysical or ethical. And if that is true, then it is possible that theoretical and practical considerations will conflict.

So, finally, in this spirit, let us contemplate the possibility that, first, a desert-based, non-consequentialist practice of responsibility is ethically speaking the one we ought to accept. But also, second, that (say) hard incompatibilists are right: such a practice is not metaphysically possible. If both of these claims were right, responsibility would be practically justified, while metaphysically impossible. The Wiggins-conjecture would then not only mark out a theoretical option, but would in fact stand for what is the case. To repeat, I have not argued above that this *is* the case, only that it *could* be the case. Even the mere possibility is significant, however, for how we should think about justifying moral responsibility.

This idea builds on but is of course already far from the original arguments put forward by Strawson. Similarly to neo-Strawsonians, it abandons the idea that inescapability could be invoked as a justificatory consideration. It goes further than neo-Strawsonians, however, in that it denies the priority of normative considerations in justifying the practice of responsibility. Yet this suggestion still retains a Strawsonian flavour insofar as it affirms the ineliminability of normative considerations from the justificatory project. I propose that this, after all, may be the most one can make of the idea of inescapability in this area. What is inescapable then is the threat of a conflict of theoretical and practical considerations in the justification of moral responsibility. What we cannot escape from is the ongoing challenge posed by the lack of a pre-established harmony between the ethics and metaphysics of responsibility.

Affiliation: Department of Philosophy, Lund University (Sweden)/
Department of Philosophy, Central European University (Hungary)
Email1: andras.szigeti@fil.lu.se
Email: szigetia@ceu.hu

REFERENCES

Ayer, A.J., 1980, "Free-Will and Rationality", In *Philosophical Subjects,* 1-13, Z. van Straaten (ed.), Oxford University Press, Oxford.

Baier, A., 1995, "Moralism and Cruelty: Reflections on Hume and Kant", In *Moral Prejudices: Essays on Ethics,* 268-293. Cambridge, MA: Harvard University Press.

Bennett, J., 1980, "Accountability", In *Philosophical Subjects,* 14-47, Z. van Straaten, Oxford University Press, Oxford.

Dennett, D., 1984, *Elbow Room,* MIT Press, Cambridge, MA.

Gosepath, S., 2009, "Zum Ursprung der Normativität", In *Sozialphilosophie und Kritik,* 250-269, R. Forst, M. Hartmann, R. Jaeggi, and M. Saar (eds.), Suhrkamp Verlag, Frankfurt am Main.

Hieronymi, P., 2004, "The Force and Fairness of Blame", *Philosophical Perspectives,* **18,** 115-148.

Hume, D., 1978 [1739/40], *A Treatise of Human Nature*, L.A. Selby-Bigge and P.H. Nidditch (eds), Oxford University Press, Oxford.

van Inwagen, P., 2008, "How to Think about the Problem of Free Will", *Journal of Ethics,* **12,** 327-341.

Kolnai, A., 1974, "Forgiveness", *Proceedings of the Aristotelian Society,* **74,** 91-106.

Korsgaard, C., 1996, "Creating the Kingdom of Ends: Reciprocity and Responsibility in Personal Relations", In *Creating the Kingdom of Ends,* 188-221, Cambridge University Press, . Cambridge.

Pereboom, D., 2007, "Hard Incompatibilism", In *Four Views on Free Will,* J. M. Fischer, R. Kane, D. Pereboom, and M. Vargas, (ed), 85-125, Blackwell, Oxford.

Rawls, J., 1971, *A Theory of Justice*, Oxford University Press, Oxford.

Russell, P., 1992, "Strawson's Way of Naturalizing Responsibility", *Ethics,* **102,** 287-302.

Schlick, M., 1962 [1939], *Problems of Ethics,* Dover, New York.

Sher, G., 2006, *In Praise of Blame*, Oxford University Press, Oxford.

Smilansky, S., 2000, *Free Will and Illusio,*. Clarendon Press, Oxford.

Stern, L., 1974, "Freedom, Blame, and Moral Community", *Journal of Philosophy,* **71,** 72-84.

Strawson, G., 1986. *Freedom and Belief,* Clarendon Press, Oxford.

Strawson, P.F., 1974, "Freedom and Resentment", In *Freedom and Resentment,* 1-25, Methuen, London.

Strawson, P.F. 1985. *Skepticism and Naturalism. Some Varieties.* London: Methuen.

Vargas, M., 2004, "Responsibility and the Aims of Theory: Strawson and Revisionism", *Pacific Philosophical Quarterly,* **85,** 218–241.

Wallace, J. R., 1994, *Responsibility and the Moral Sentiments,* Harvard University Press, Cambridge, MA.

Wallace, J. R., 2006, "Moral Responsibility and the Practical Point of View", In *Normativity and the Will,* 144-64, Clarendon Press, Oxford.

Watson, G., 1987, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme", In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology,* F. Schoeman (ed.), 256-286, Cambridge University Press, Cambridge.

Wertheimer, R., 1998, "Constraining Condemning", *Ethics,* **108,** 489-501.

Wiggins, D., 2002, "Towards a Reasonable Libertarianism", In *Needs, Values, Truth,* 269-302, 3rd revised edition, Oxford University Press, Oxford.

Williams, B., 1976, "Moral Luck", In *Proceedings of the Aristotelian Society,* suppl. vol. **L,** 115-135.

Williams, B., 1993, *Shame and Necessity,* University of California Press, Berkeley.

Williams, M., 1996, *Unnatural Doubts,* Princeton University Press, Princeton.

Wolf, S., 1981, "The Importance of Free Will", *Mind,* **60,** 386-405.