Putting Psychology Before Metaphysics in the Arena of Moral Responsibility: What Cognitive Neuroscience Tells Us About the Reactive Attitudes

Abstract

P.F. Strawson argues that the practices of holding people morally responsible stem from our nature as social beings and have a justificatory status that is independent of whatever beliefs we may have about the metaphysical conditions necessary for morally responsible action. Using some recent work from cognitive neuroscience, we develop something lacking from Strawson’s original discussion: an empirically informed account of how the reactive attitudes flow from our nature. While this account does not vindicate Strawson’s claim that the reactive attitudes are psychologically unavoidable, it does support a related claim: that it is psychologically impossible to avoid the gut feeling, produced by a cognitive system dubbed “the Person Representation Network”, that certain entities are persons and therefore appropriate targets for the reactive attitudes. Building off the latter claim, we arrive at an account of moral responsibility that, like Strawson’s original account, puts psychology before metaphysics but which also avoids some of the shortcomings of Strawson’s account.

Introduction

Peter Strawson’s “Freedom and Responsibility” presents a number of important challenges to the philosophical literature on moral responsibility. In this paper, we focus on just one of the many influential lines of argument from that piece. (This is not to say that this argument is superior to, or more central than, any of the other arguments present in “Freedom and Responsibility”.) This argument challenges the assumption that our practices of holding people morally responsible for their actions are driven by beliefs that we have about their having satisfied the metaphysical conditions necessary for their actions to qualify as being free.¹ In challenging this assumption, Strawson makes several descriptive claims about where (and how) the practices of holding one another morally responsible spring from our mental economy. These descriptive claims, in turn, lead him to develop an account of the justificatory status of those practices that relegates the kinds of metaphysical issues that dominate much of the literature on this topic to the sidelines.

¹ People who make this assumption often claim that the beliefs in question are pre-reflective in the philosophically untutored; it is the job of the philosopher, in part, to uncover and explicitly identify the metaphysical conditions to which such beliefs are sensitive; our contention, like Strawson’s, is that the number of contexts that provoke such beliefs, even prereflectively, is vastly overestimated.
We are interested in the general approach to moral responsibility embodied by this line of argument from Strawson: we are interested in an approach that begins with descriptive claims about the psychological source of our practices of holding people moral responsible and terminates in a position where the justificatory status of those practices is, in a sense, independent of any beliefs we may have about the metaphysics of morally responsible action. Let’s call this general kind of approach to moral responsibility a “Psychology-Before-Metaphysics” or “PBM” approach. In what follows, we review a line of argument from “Freedom and Responsibility” that details a PBM approach to moral responsibility (section 1). After leveling some concerns against the particulars of this argument (section 2), we develop an alternative PBM approach that uses resources from cognitive neuroscience to give an expanded account of how the reactive attitudes flow from our nature as social beings (sections 3 and 4). This, in turn, will lead us to another PBM account of moral responsibility, a version that we think has certain advantages over Strawson’s original version.

1. Strawson’s version of the PBM approach

Philosophical work on moral responsibility typically assumes that certain metaphysical conditions must obtain in order for an action to be truly free and, thus, for an agent to be morally responsible for that action. As a result, much time and effort has been spent trying to articulate what these metaphysical conditions are and whether they are compatible with determinism, god’s omniscience, quantum-level indeterminacies, and

---

2 When Strawson discusses the practices of holding people morally responsible for their actions, he often slides between framing the issue in terms of whether these practices are justified and framing it in terms of whether they are rational. For expositional simplicity, we will focus (mostly) on justification in what follows.
so on. In “Freedom and Resentment”, P.F. Strawson presents several arguments for thinking that this infatuation with the metaphysical conditions necessary for morally responsible action over-intellectualizes the real-world phenomenon of holding one another responsible. In the place of this metaphysics-centered approach, he offers an alternative approach to moral responsibility that he suggests rests upon a more psychologically realistic picture of how we actually go about holding one another morally responsible. In this rival approach, the kinds of metaphysical questions that occupy a place of prominence in most philosophical accounts of moral responsibility end up being relegated to the sidelines because, in a variety of ways, the activities of holding one another morally responsible are independent of such metaphysical questions and issues.

Strawson begins his discussion by claiming that the center of our morality involves a set of practices built around “reactive attitudes” that are connected to the maintaining of our interpersonal relationships. Attitudes like resentment, indignation, hurt feelings, anger, gratitude, reciprocal love, and forgiveness are reactions to the degree to which we feel that others are acting towards us with a good will, an ill will, or indifference. (The “moral attitudes”, in turn, are simply generalized or vicarious versions of the reactive attitudes; they are reactive attitudes that reflect the degree to which we feel that someone is acting with a good will, an ill will, or indifference towards anybody, and not just towards oneself. In our examination of the reactive attitudes, we are going to

---

3 Many of these arguments are continued in Strawson 1985. For good overviews of the additional arguments from Strawson in favor of this alternative approach to moral responsibility, see Eshleman 2004 and McKenna and Russell 2008.

4 Understood as a kind of reactive attitude, the activities of holding one another morally responsible involves something more than the desire to shape the character or influence the future behaviors of one another. The point is also to react to another person in a way that reflects what she deserves in virtue of the will she has exhibited towards you.
focus on a particular argument from Strawson that begins with a descriptive claim about our psychology. The claim is that the reactive attitudes flow from something within our nature as social beings, and not from our having any beliefs about the metaphysical workings of our conspecifics; they flow from “our natural human commitment to ordinary interpersonal relationships” (p. 27-28) and have “roots in our human nature and our membership of human communities” (p. 29). To express the idea metaphorically, we come to the philosopher’s table — the table where issues and questions about the metaphysical conditions necessary for free (and morally responsible) action are raised and discussed — possessing social natures that already lead us, via the reactive attitudes, to hold people morally responsible for their actions.

Strawson says very little about what it is within our nature as social creatures that causes these attitudes; he states that he is “not much concerned” (p. 23) with the question of the general causes of the reactive attitudes. Despite not saying anything about where these attitudes spring from, besides that they stem from something within us connected to our nature as social beings, Strawson does make a second descriptive claim about how the reactive attitudes flow from our psychology: specifically, that we are incapable of permanently withdrawing from the reactive attitudes. Regardless of what theoretical conclusion we may reach at the philosopher’s table about the metaphysical conditions necessary for free action, the reactive attitudes will continue to arise in us whether we want them to or not.

A sustained objectivity of inter-personal attitude, and the human isolation which that would entail, does not seem to be something of which human beings would be capable, even if some general truth were a theoretical ground for it. (p. 26-27)

The commitment to participation in ordinary inter-personal relationships is, I think, too thoroughgoing and deeply rooted for us to take seriously the thought that a general theoretical conviction might so change our world that, in it, there were no longer any such things as inter-personal relationships as we normally understand them… (p. 26)
Strawson allows that we can temporarily refrain from the reactive attitudes if we believe that certain excusing conditions obtain — if, for instance, you believe a person’s behavior was an accident or unintentional and therefore not a reflection of her will towards you. We can also permanently exempt a particular person from these attitudes altogether — if, for example, you believe a subject is cognitively impaired or does not have a fully formed and functioning will. Strawson even grants that, with effort, we can temporarily refrain from adopting the reactive attitudes towards someone who doesn’t satisfy any of the excusing or exempting conditions listed above; simply as a matter of convenience, you can temporarily adopt an “objective” attitude towards a subject and view her in purely causal terms, as an object for treatment or control, and not as an entity with a will deserving of resentment or praise. (Anyone required to engage in administrative meetings within a university setting is likely to be familiar with this strategy for managing people.)

What we can’t do, according to Strawson, is turn the reactive attitudes completely off. Strawson makes this claim several times in “Freedom and Resentment” and constructs a clear line of argument around it — a line of argument that we are currently examining. (Whether or not he “really” believes this claim is perhaps debatable.⁵) If this claim is true, then the reactive attitudes cannot be permanently brought offline by any theoretical conclusions reached at the philosopher’s table about whether or not people generally satisfy the metaphysical conditions necessary for free action.

⁵ The problem is that there are other passages in “Freedom and Resentment” where Strawson backtracks on this claim and seemingly concedes that it may, in fact, be possible to permanently shut off the reactive attitudes. In these parts of “Freedom and Resentment”, Strawson offers other arguments for thinking that metaphysics should be relegated to the sidelines in our discussion of moral responsibility. Although they are fascinating, we will not be considering any of these other arguments in this paper.
Strawson doesn’t say much about the exact sense in which it is “impossible” for us to take the reactive attitudes permanently offline. Certainly it’s not *logically* impossible to live a life where these attitudes never happen to arise: in a number of ways an all-powerful evil demon could make you to live a life where you never adopted the reactive attitudes towards anything. It’s also probably not *physically* impossible for these attitudes to be permanently shut off; given the right kind of brain damage, it seems likely that these attitudes would never arise within a subject. Given these points, we submit that the most plausible of Strawson’s claim is that it is impossible for someone possessing a “normal” brain, existing in “normal” circumstances – that include regular confrontations with other human beings – o never undergo the reactive attitudes. We’ll express this idea by saying that it is “psychologically” (as opposed to “logically” or “physically”) impossible to avoid the reactive attitudes: given a normal subject’s psychology, placed in normal situations, and regularly interacting with other people, the reactive attitudes cannot be permanently avoided.

This second descriptive claim leads to an argument concerning the justificatory status of the reactive attitudes. If it was within our power to choose whether, in general, we would hold one another morally responsible for our actions, then it would make sense to ask whether or not the choice in favor of such a practice is justified or not. In asking this question, we would be looking for some kind of external justification in favor of this choice — i.e. a justification of the practice of holding people responsible that is *external* to that practice. And one place we might turn in an effort to find such external justification would be metaphysics, in particular, theoretical reasoning concerning the metaphysical conditions that are necessary for morally responsible actions.
Strawson line of argument gets off this ride at the first stop: it *rejects* the assumption that there is a genuine choice to be made about whether, in general, we will hold people morally responsible for their actions. Strawson maintains that such behaviors are underpinned by the reactive attitudes, reactions that flow directly from our nature as social beings that cannot be turned off (short of the interference of an evil demon or brain damage). Since there is no real “choice”, Strawson claims that it doesn’t make sense to ask whether these behaviors are justified or not; as he puts it, it is “useless” to ask whether it is irrational to do what, given our nature, we cannot avoid doing.\(^6\)

The result of this line of thought is a Compatibilism of sorts. It’s Compatibilist in the sense that it rejects the claim that determinism and morally responsible action are incompatible. It differs from more prototypical Compatibilist positions, however, in that it doesn’t try to demonstrate that the metaphysical conditions necessary for free action are compatible with determinism by giving an analysis of the former. Instead, it argues that the justificatory status of the practices of holding one another moral responsible are *insulated* from such metaphysical questions—the truth or falsity of a thesis like determinism is irrelevant to the descriptive question of whether we will continue hold people responsible and the normative question of whether we are justified in doing so.

\(^6\) Strawson allows that particular instances of the reactive attitudes — instances of resentment, indignation, forgiveness, etc. — can be described as justified or unjustified because from within the system of reactive attitudes there are rules that govern when a given reactive attitude is appropriate and when it is not. What we are talking about, however, is the justificatory status of the reactive attitudes *considered as a whole* and from a perspective external to that practice.
2. Is Strawson’s naturalism naturalistic enough?

   Earlier, we characterized a PBM approach to moral responsibility as being an approach that begins with claims about the psychological source of the practices of holding one another morally responsible for our actions and, on the basis of those descriptive claims, reaches a position where the justificatory status of those practices is in some sense independent of theoretical reasoning about metaphysics. The line of argument we just reviewed from Strawson makes such an approach, but there are other ways of filling in the details of a PBM approach. We are interested in developing a PBM approach that uses the resources of cognitive neuroscience to generate a more detailed account of how it is that the reactive attitudes flow from our nature as social beings. Although Strawson didn’t seem to think that such an appeal would add much to the discussion, we think it could pay dividends in a number of ways. We applaud Strawson’s naturalism in undertaking the topic of moral responsibility, but we think it isn’t quite naturalistic enough.

   To get a sense of some of the ways in which an empirical account of how the reactive attitudes flow from our social nature could pay dividends, let’s consider some of the objections that have been leveled against Strawson’s version of the PBM approach. The following is not meant to be a complete catalogue of all the objections that have been leveled against Strawson’s argument; instead, these are merely examples of objections that demand a more detailed empirical account of how the reactive attitudes stem from our nature. (They also foreshadow some ways in which such details could result in a version of a PBM account that differs from that given by Strawson.)
To start with, it’s easy to generate doubt about the claim at the heart of Strawson’s argument concerning the justificatory status of the reactive attitudes, the claim that it is (psychologically) impossible to completely avoid the reactive attitudes. Indeed, there are things that Strawson himself says which invite such doubt. Recall that Strawson allows that it’s psychologically possible to temporarily refrain from the reactive attitudes, in a variety of ways. This leads to a problematic question for Strawson: if the psychological mechanisms responsible for the generation of the reactive attitudes can be overridden some of the time, why can’t they be overridden all of the time? Since he is “not much concerned” to detail the empirical causes of the reactive attitudes, this entire line of argument appears to be precariously balanced upon a claim about the inner causes of the reactive attitudes — namely, that they are produced in a manner that makes them psychologically impossible to avoid. This, however, is a claim that Strawson takes no interest in empirically investigating or defending. Obviously, an empirical account of these mechanisms could yield resources that speak to this doubt. Even if it didn’t, it could uncover others ways in which these mechanisms – the ways these attitudes flow from our nature – have implications for the justificatory status of these attitudes. The justifications discovered here might be distinct from Strawson’s claims about the psychological impossibility of permanently shutting them off.

Here’s another example: some have objected to Strawson on the grounds that the core idea of there being an incompatibility between determinism and moral responsibility comes from just as “deep” a place within our nature as the that from which the reactive attitudes arise. This raises the possibility that whatever special epistemic status the
reactive attitudes enjoy could also be enjoyed by the belief in Incompatibilism. An empirical account of how the reactive attitudes arise from our nature has the potential to reveal significant differences between how our nature gives rise to the reactive attitudes and how it gives rise to the core idea of incompatibility, which, in turn, could explain why reactive attitudes have some of the special justificatory properties that Strawson attributes to them, while the latter, perhaps, do not.

A final example of an objection that calls out for an empirical account of the how the reactive attitudes flow from our nature as social beings involves the claim that the reactive attitudes are more flexible—they are more a matter of culture—than Strawson makes them out to be. An empirically informed account of how our nature gives rise to these attitudes could provide the resources necessary to uncover the extent to which they are, and are not, capable of being shaped by culture.

In the next section, we undertake the project of using the empirical sciences to provide a more detailed account of how the reactive attitudes “flow from our nature as social beings” and not from whatever beliefs we have about the metaphysical inner-workings of our conspecifics. Towards this end, we’ll explore some recent work from cognitive neuroscience concerning a neural network that appears to be implicated in the production of the reactive attitudes.

3. The Person Representational Network and the reactive attitudes

---

7 For an instance of this objection, see G. Strawson 1986. Nichols and Knobe (2007) explore the question of whether the Incompatibilist belief is “fundamental” using X-phi methodology. They do not, however, give a detailed empirical account of how it flows from our nature.
8 See, for instance, Pereboom 2001.
Our empirical account of how the reactive attitudes flow from our nature is built around Martha Farah and Andrea Heberlein’s (2007a, 2007b) description of an innate, autonomous cognitive network that they dub the “Person Representation Network” (PRN). Although the existence of this neural network is a substantial empirical thesis and thus susceptible to refutation via new data, experiments, etc., we will be taking the existence of PRN largely for granted. Our focus will be on determining how Farah and Heberlein’s description of this cognitive network can be used to illuminate the claim that the reactive attitudes “flow from our nature as social beings”.

PRN is a cognitive network of systems that, when triggered, generates a gut feeling that certain entities in the surrounding environment are persons. We will treat this feeling as being both phenomenal and representational: it’s phenomenal in that there is “something that it’s like” to undergo the feeling and it’s representational in that the feeling carries the representational content that some object in the surrounding environment is a person. In virtue of representing the latter, the feeling represents an object as being the kind of thing capable of entering into interpersonal relationships with the subject — it represents it as being the kind of thing capable of acting towards us with a will that may be good, ill, or indifferent. In short, it represents the object in question as being a “person” in the moral sense of the term.10

---

9 This network has also been referred to as “the social brain”. See, for instance, Brothers 1990, Adolphs 2003, and Skuse et al. 2003.
10 For the record, Farah and Heberlein do not explicitly mention the reactive attitudes when describing the gut feelings that are the output of PRN. Given that their interests reside in the understanding how the operations of PRN lead to the conception of a “person” where this is understood as a moral concept, however, we think our claims about the representational content of this feeling are appropriate, given the Strawsonian framework from which we are working.
If this feeling is representational, then misrepresentation should be possible, and indeed it is. As you’ll see shortly, PRN will represent the presence of a person in virtue of being triggered by any number of diverse properties and behaviors, including some properties and behaviors that are only superficially correlated with the presence of actual persons. This means that PRN is a fallible mechanism that can generate false positives: it can misrepresent the presence of persons in a wide range of cases where no actual persons are present. This apparent weakness in design makes sense from an evolutionary perspective. Our ancestors were social creatures that were reliant upon one another in a variety of ways, and the ability to quickly identify entities as being the kinds of things with which they could socially interact was linked to their survival. Given this, it was far better for our evolutionary ancestors (and their brains) to err on the side of attributing personhood to things that weren’t persons than it was to make the mistake of not attributing personhood to things that were, in fact, persons.

As it indicates that something is a person and, hence, an appropriate target for the reactive attitudes, the content of this feeling is neutral on the internal metaphysics of that thing; the gut feeling “says” that the object in question is a person while not saying anything about whether the internal behavior-generating-mechanisms of that object are governed by deterministic causal law or not. Also, though it represents that some object in the surrounding environment is a person — it is an entity capable of socially interacting with us with a good will, an ill will, or an indifferent will — the gut feeling generated by PRN does not indicate whether that entity’s will is, in fact, good, ill, or indifferent. Some other cognitive system, or systems, performs that task. For this reason, the gut feeling generated by PRN does not, itself, count as a reactive attitude. Instead, we
view it as a *causal precursor* to the reactive attitudes: it is something that represents an object as being an appropriate target for such attitudes and, as a result, can cause us to adopt the reactive attitudes towards that object.

Let’s temporarily move away from the gut feeling produced by this cognitive network and take a closer look at some of the cognitive subsystems that make up PRN. According to Farah and Heberlein, this network is composed of a host of neural mechanisms that specialize in identifying our conspecifics and facilitating social interaction with them, including:

- A region of the fusiform gyrus that specializes in visually recognizing human faces. This region is not active when recognizing other kinds of complex shapes/objects.
- Another part of the fusiform gyrus and the lateral surface of the brain near the temporoparietal juncture that specialize in classifying certain complex shapes (e.g. stick figures) as human bodies. These regions are not active during the perception of other kinds of equally complex shapes.
- A region on the lateral surface of the brain near the temporoparietal juncture that specializes in registering actions/movements that are similar to those of human bodies. This region is not active during the perception of the movements of other kinds of objects.
- A region in medial prefrontal cortex that is active when listening to stories that involve understanding another subject’s mental states. This area is not active when listening to equally complex stories that just require the subject to understand/think about complex physical causation.

Many of the mechanisms making up PRN are hardwired, as evidenced by their operation in the non-linguistic behavior of infants. For example:

- Shapes that move in mutually contingent patterns relative to one another (i.e. patterns that could be described as involving those shapes “reacting to one another”) hold the attention of 3 month-old infants longer than shapes that move in non-contingent ways.  
11
- 12 month-old infants follow the gaze of objects that are shaped like eyes — in short, they look where these “eyes” seem to be looking — as long as those eye-

---

11 For additional details, see Rochat et al. 1997.
shaped objects are moving in a contingent manner.\textsuperscript{12}

As a result of containing these and other hardwired mechanisms, the gut feeling generated by PRN is, in an important sense, \textit{pre-theoretical}: long before we take philosophy classes and consider various candidates for the metaphysical conditions necessary for morally responsible action, our brains are already issuing forth gut feelings that certain entities are persons and, hence, (qua our view) appropriate targets for certain attitudes including the reactive attitudes. Because of the hardwired operations of PRN, we come to the philosophers’ table already in possession of gut feelings that, as a general rule, there are certain entities in our environment to which the reactive attitudes are appropriate responses.

Not only is PRN already generating gut feelings that certain entities are appropriate targets for the reactive attitudes before we undertake any philosophical theorizing about the metaphysics of moral responsibility, it will continue to generate those feelings after we undertake such theorizing. PRN generates these feelings in a way that is indifferent to such theorizing. More carefully, the operation of PRN is \textit{autonomous} from such theorizing; in fact, PRN appears to be autonomous from all other network/systems of the brain apart from the perceptual systems that provide it with its inputs. The autonomy of the PRN can be put on display in a number of scenarios where PRN generates the relevant gut feelings despite the subject knowing that he or she is not, in fact, dealing with a person. For example:

- The presence of cartoon eyes, which are consciously recognized by subjects as such, leads subjects to adopt more generous strategies in computer-run economics games.\textsuperscript{13}

\textsuperscript{12} For additional details, see Johnson 2003.
\textsuperscript{13} For additional details, see Rochat et al. 1997.
• Subjects cannot help but experience the moving geometrical shapes in Heider and Simmel’s (1944) animated film — a film that involves one large triangle moving around a space with two smaller figures, a triangle and a circle — as having motivations; they cannot help but experience the large triangle as pursuing and attacking the other two “terrified” shapes.

Because of this autonomy, the gut feelings generated by PRN will persist in the face of whatever information the rest of the brain has about the entity in question. So regardless of what we learn about the inner workings of ourselves and of our conspecifics, and regardless of what our philosophical theories tell us about the justificatory status of holding people responsible for their actions, PRN will continue to generate gut feelings that certain entities are appropriate targets for these reactive attitudes.

4. A second PBM account of moral responsibility

The preceding discussion of PRN accommodates the descriptive claim, made by Strawson, that the reactive attitudes stem from something within us connected to our social nature and not from whatever beliefs we may have about the metaphysical inner workings of our conspecifics. The gut feeling we described flows from PRN — a cognitive network that specializes in quick, heuristically driven, identifications of person-like entities for purposes of facilitating social interactions. As we’ve seen, PRN identifies certain entities as being appropriate targets for the reactive attitudes pre-theoretically: we come to the table where philosophers fret about the metaphysical conditions necessary for morally responsible action already having gut feelings that some entities are appropriate targets for these attitudes.

Let’s now turn to Strawson’s claim that the reactive attitudes flow from our nature in a manner that makes them psychologically impossible to resist. Does the previous
discussion of PRN and the gut feelings it produces also support this descriptive claim? As we’ve seen, PRN operates autonomously from much of the rest of the brain; it will continue to generate the relevant gut feelings, given the appropriate perceptual triggers, regardless of any theoretical conclusions we may reach about the metaphysical conditions necessary for a subject to truly responsible for her actions. So, barring the interference of an all-powerful evil demon or brain damage to the neural mechanisms composing PRN, it is psychologically impossible to avoid the gut feelings that PRN produces when it is triggered by various stimuli.

But this doesn’t mean that the reactive attitudes will also be psychologically impossible to avoid. Recall that these gut feelings are causal intermediaries to the reactive attitudes; they are something that, when they occur, can cause the subject to adopt the reactive attitudes towards the entity in question. For all that we’ve said, it may be psychologically possible for a subject to experience the gut feelings generated by PRN while never undergoing the reactive attitudes; it may be that these gut feelings dispose (or strongly dispose) the subject towards the reactive attitudes and, despite that fact, it’s psychologically possible (perhaps with substantial effort) to permanently override this disposition. So despite uncovering something that is causally related to the reactive attitudes — the gut feeling generated by PRN — that is psychologically impossible to avoid, our empirical account has not vindicated Strawson’s claim that the reactive attitudes themselves are psychologically impossible to permanently avoid.

Is there some other way in which our discussion of PRN and the gut feelings it produces could support to a PBM account where the justificatory status of the reactive attitudes is, in some sense, independent of our beliefs about the metaphysical inner-
workings of our conspecifics? We believe there is. The first step is to adopt a foundationalist epistemology that maintains that perceptual experiences serve as immediate (but defeasible) justification for beliefs about the surrounding environment. More specifically, we want to adopt a version of this position that maintains that it is the perceptual phenomenology of these states that give them their justificatory powers: perceptual phenomenal character serves as a form of immediate, but defeasible, justification for beliefs that appropriately mirror the representational contents of these phenomenal states.\textsuperscript{14} According to such a position, the phenomenal character of my visual experience as of the chicken in the yard provides me with immediate, but defeasible, justification for the belief that there is a chicken in the yard.

Given this version of foundationalism, the feeling generated by PRN provides immediate, but defeasible, justification for the belief that some entity before the eyes is a person and thus an appropriate target for the reactive attitudes. Because this is a type of foundational justification, a subject does not need not possess justified beliefs about anything else — including justified beliefs about the metaphysical conditions necessary for morally responsible action — in order to possess this justification for believing they are in the presence of a person who is morally responsible for her actions. In this manner, our empirical investigation into PRN has revealed another way in which the reactive attitudes have a justificatory status that is, in a sense, \textit{independent} of any beliefs we may have about the metaphysical inner workings of our conspecifics.

It’s worth taking a moment to highlight, in more detail, how our PBM account differs from Strawson’s on this front. According to Strawson’s PBM account, there is no

\textsuperscript{14} For good discussion, and defenses, of such a position, see, for example, Pryor 2000 and Huemer 2007.
real choice about whether, in general, we should adopt the reactive attitudes towards one another because the reactive attitudes are generated in a manner that makes them psychologically impossible to avoid. For this reason, Strawson asserts that the question of whether or not these attitudes can be justified via theoretical reasoning about the metaphysics of morally responsible action is misplaced; it’s “useless” to ask whether it’s rational to do what, given our nature, we cannot avoid doing. In contrast, our (more empirical) account of the reactive attitudes allows that there could be a choice about whether, in general, we should adopt these attitudes or not; although the gut feeling produced by PRN is psychologically impossible to avoid, it’s an open question whether this feeling will make the reactive attitudes causally inevitable. This means that, unlike Strawson, we allow that there may be a need to justify the “choice” of holding people responsible. On our view, however, there is a kind of justification available for making this choice that has nothing to do with what takes place at the philosopher’s table; instead, it involves the gut feelings generated by PRN. Since these feelings provide defeasible justification for the reactive attitudes, they can serve as a source of justification for the adoption of those attitudes that is independent of any theorizing about the metaphysical conditions necessary for morally responsible action.

Another significant difference between these two PBM accounts concerns the ability of theoretical reasoning about the metaphysical conditions necessary for morally responsible action to undermine one’s justification for adopting the reactive attitudes towards one another. Strawson argues that since it is psychologically impossible to permanently refrain from the reactive attitudes, these attitudes cannot be rendered irrational by theoretical reasoning that is external to the practice. Our account, in
contrast, is consistent with the possibility that there is a choice about whether or not to adopt the reactive attitudes. Although we’ve argued that the choice in favor of the reactive attitudes can be justified by a process that is independent of any theoretical reasoning about determinism and its implications — it can be justified by the justificatory powers of the gut feeling generated by PRN — there is nothing in our account that secures the claim that the justificatory powers of the feeling cannot be defeated by theoretical reasoning about metaphysics. Recall that the phenomenology of these representational gut feelings provides *prima facie*, and thus defeasible, justification for one’s beliefs. So while there is a source of justification for adopting the reactive attitudes that is epistemically isolated from any theoretical reasoning about determinism, we do not claim that the legitimacy of these attitudes is completely immune to defeat from such reasoning.

This last difference between our position and that of Strawson’s may stick in the craw of some; they might complain “What’s the point of developing a PBM account of moral responsibility if you cannot secure a position, like Strawson’s, that *completely* *insulates* the justificatory status of the reactive attitudes relative to philosophical theorizing about the metaphysical conditions necessary for moral responsibility?” There are a couple of points to make in response. First, it’s not clear that Strawson is really entitled to make such claim in the first place. As we’ve seen, his argument in favor of it relies upon a descriptive claim — the claim that the reactive attitudes are psychologically impossible to avoid — that is empirically under-supported. In addition, even if this descriptive claim were true, that wouldn’t guarantee that the justificatory status of the reactive attitudes is completely immune to theoretical reasoning about the metaphysical
conditions necessary for morally responsible action; someone might infer, instead, that
the proper conclusion to draw is that we should live with the cognitive dissonance of
being committed, via our nature, to a practice that we know is unjustified.

Second, even though our PBM approach doesn’t completely insulate the
justificatory status of the reactive attitudes from theoretical reasoning concerning the
metaphysical conditions necessary for morally responsible action, it does uncover
something significant about the burden proof in these debates. What’s more, it does so in
a way that results in a Compatibilist position that is somewhat atypical in a manner akin
to Strawson’s Compatibilist position. For neither our account nor Strawson’s argues for
the compatibility for moral responsibility and determinism (or any other metaphysical
thesis) by first articulating an account of the metaphysical conditions necessary for
morally responsible action and then showing that those conditions are compatible with
determinism. Like Strawson’s account, our defense of Compatibilism doesn’t say
anything about such metaphysical conditions. Instead, our account maintains that a belief
in Compatibilism is initially (and defeasibly) justified in a manner that is independent of
any philosophical investigation into the metaphysical conditions necessary for free
action, in virtue of a gut feelings that is generated by PRN.

5. Conclusion: some reasons for favoring our PBM account over Strawson’s

As PBM accounts, both our position and Strawson’s are built around descriptive
claims about our psychology having to do with how the reactive attitudes flow from our
nature as social beings. The descriptive claims we make on this front stem from
contemporary empirical research concerning PRN. The descriptive claims Strawson
makes, in contrast, seem to come solely from the armchair; he professes having no interest in investigating the inner causes of the reactive attitudes. In light of this difference, the starting point for our argument — a starting point which involves descriptive claims about our psychology — is far better supported, empirically, than the analogous starting point for Strawson’s argument.

The differences between our positions also allow us to take some new angles on the objections against Strawson’s argument that were mentioned near the end of section 2. Some deny Strawson’s claim that it is psychologically impossible to permanently refrain from the reactive attitudes; these folks claim that a general conclusion from metaphysics could, in fact, cause us to abandon the reactive attitudes altogether. Given that he has eschewed empirical investigation into the inner causes of the reactive attitudes, Strawson’s hands are tied in responding to this objection — it seems that all he can do is pound the table and insist that something (unspecified) in our nature makes completely abandoning the reactive attitudes psychologically impossible.

Our account, by contrast, does not assert that it is psychologically impossible to permanently refrain from the reactive attitudes. Instead it claims that it is psychologically impossible to avoid the gut feelings generated by PRN. By laying out Farah and Heberlein’s empirical case in favor of the latter claim, we have provided a framework that sets the agenda for exploring and adjudicating possible objections to this descriptive psychological claim. If, for instance, someone wanted to deny that these gut feelings are psychologically unavoidable, both parties involved in this debate would know where to go for the relevant evidence: namely, psychological data/experiments that speak in favor the autonomy of PRN relative to most other cognitive systems and networks.
Another objection to Strawson’s version of the PBM approach is that the basic idea of the incompatibility of determinism and moral responsibility comes from a place just as “deep” from within our nature as the commitment of the reactive attitudes, so whatever special epistemic status the reactive attitudes enjoy will also be shared by the belief in Incompatibilism. Our PBM account is in a position to accept the claim that the belief in Incompatibilism runs deep while also maintaining that there is something epistemically special about the reactive attitudes. Under our account, what gives the reactive attitudes their epistemic status isn’t the mere fact that they flow from deep within our nature. What matters is how they flow from our nature; what matters is the fact that something that produces these attitudes — the gut feelings of PRN — also provides a kind of foundational justification for them in virtue of its perceptual phenomenology. In absence of a detailed empirical account of how the basic Incompatibilist idea “flows from our nature” that connects that process to something that confers justification upon that belief, there is no reason for thinking that this idea will share the same special justificatory status as the reactive attitudes.

Finally, some assert that the reactive attitudes are heavily influenced by culture and, as a result, are less determined by our nature than Strawson makes them out to be. Our account can also accommodate this claim, to a degree. What our “social nature” guarantees is that, when confronted by entities that trigger PRN, we will have gut feelings that we are confronting persons and, as a result, be disposed towards reactions that reflect what we take their wills towards us (and others) to be. This position is consistent, however, with there being considerable flexibility in terms of the precise nature of these reactions to one another’s will. (Recall that our position is that the gut feeling from PRN
pushes us towards the reactive attitudes but that other cognitive systems determine the
form that those reactive attitudes take.) To put it another way, while the gut feeling that
something is acting with a will towards you is “universal” — everyone with a functioning
PRN and in normal circumstances will experience it — the precise nature of your
reaction towards entities that trigger this feeling could be determined largely by culture.

References


