Strawsonian Moral Responsibility, Response-dependence, and the Possibility of Global Error

Abstract

Various philosophers have wanted to move from a (P.F.) "Strawsonian" understanding of the "practices of moral responsibility" to a non-skeptical result. I focus on a strategy moving from a "response-dependent" theory of responsibility. I aim to show that a key analogy associated with this strategy fails to support a compatibilist result. It seems clear that nothing could show that *nothing we have been laughing at has really been funny.* If "the funny" is similar to "the blameworthy", then perhaps it would follow that nothing could show that *no one we have ever blamed has really been blameworthy.* The comparison is interesting, but inconclusive: even if we grant that these properties are normatively similar, the latter has substantive empirical presuppositions, whereas the former does not. One important upshot: even if the standards operative in our practices cannot be mistaken, it could nevertheless be that no one has ever met these standards.

Philosophers have often maintained that P.F. Strawson's "Freedom and Resentment" contains a radical kind of reorientation of the traditional understanding of moral responsibility, a reorientation somehow pertaining to the relative priority of *being responsible* and *holding responsible.* This thesis – sometimes called Strawson's "reversal" thesis – has recently come to be associated with a so-called *response-dependent* theory of moral responsibility, a theory of responsibility notably defended at length in several recent works by David Shoemaker (2017, 2022). According to Shoemaker, it is only when we understand the import of Strawson's "response-dependent" conception of responsibility that we can appreciate Strawson's central line of argument against the "pessimistic" incompatibilist. Indeed, for Shoemaker, the response-dependence thesis is crucial to understanding Strawson's famous insistence that it is a mistake to seek any kind of "external justification" of our responsibility practices – a mistake Strawson (and Shoemaker) see lurking behind incompatibilist argumentation.

In this paper, I first contend that the response-dependence thesis articulated by Shoemaker provides no response to the incompatibilist. The response-dependence thesis may (or may not) be plausible, but that thesis has little or nothing to do with the traditional debate between the compatibilist and the incompatibilist. Ultimately, then, even if the thesis is plausible, it cannot play the role Strawsonians have wanted it to play – viz., advancing Strawson's key compatibilist argumentative agenda in "Freedom and Resentment". Broadly speaking, Shoemaker's development of this theory faces a key dilemma: either the theory is ultimately *normative* or it isn't. If it is, then it won't entail compatibilism – and indeed, it will leave any such questions exactly as they were. If it isn't, however, then (a) it is implausible on other grounds, a point Shoemaker is perfectly happy to concede, and (b) as I hope to show below, it may not support compatibilism anyway.

My purpose here, however, is not limited solely to making this criticism of Shoemaker's project. More generally, I wish to show why no plausible response-dependence thesis can in principle decide the debate between the compatibilist and the incompatibilist. To accomplish this aim, I turn to an analogy first introduced by Todd (2016), and which subsequently structures much of Shoemaker's discussion – an analogy between *the blameworthy* and *the funny*. Todd observed the following. Prima facie, it seems strange to think that it could turn out that nothing we have ever laughed at has been funny. In some hard to specify sense, we might think, our dispositions to amusement somehow *fix* what is funny and what isn't – and thus it couldn't turn out that our dispositions to amusement are simply *entirely* on the wrong track. But if "the funny" is, in this respect, strongly analogous to "the blameworthy", then perhaps it similarly could not turn out that no one we've ever blamed has been blameworthy. In this paper, I aim to uncover why this style of argument fails, *even if the response-dependence thesis is granted*. The incompatibilist can grant that our dispositions to blame – suitably described – fix the standards of blameworthiness. However, it nevertheless remains a substantive question *whether anyone has in fact met those standards*. The key difference between the funny and the blameworthy thus emerges: the latter, the incompatibilist will say, has substantive empirical presuppositions, whereas the former does not.

Before beginning, let me clear the air. Strawson's original essay has been interpreted in various different ways; indeed, how best to interpret that essay is an ongoing debate amongst theorists of responsibility. The purpose of this paper, however, is not exegetical; I make no claims regarding whether Strawson did or did not endorse anything like the "response-dependence" thesis discussed in this paper. (However, the discussion below will certainly shed light on this interpretive question.) Further, just as Strawson's whole essay has been interpreted in various ways, the "reversal" thesis in question has also been developed in various differing ways. For instance, David Beglin, and Benjamin De Mesel and Sybren Heyndels have recently offered interpretations of this "reversal" which engage many of the themes discussed below. However, for these theorists, the "reversal" in question simply *sets the proper stage* on which the debate between the incompatibilist and compatibilist should proceed; unlike Shoemaker, they simply admit that the relevant thesis is consistent with both incompatibilist and compatibilist views.[1] (However, Beglin contends that the thesis somehow *favors* compatibilist views. As we will see below, I disagree.) In the following, I thus focus mostly closely on Shoemaker's presentation, where the following theme will emerge: once we go slowly, and unpack precisely what the response-dependence claim really is, it becomes clear that the thesis cannot support the argumentative burden Shoemaker requires of it.

Let me clear the air yet further. The terminology of "response-dependence" has been invoked in a wide variety of philosophical settings, and distinctions have been drawn in these settings to attempt to capture the diversity of the phenomena at stake. I thus wish to flag in advance that the terminology of "response-dependence" at this point seems fundamentally contested.[2] I will thus for the most part bypass the question of whether a given analysis deserves the title of a "response-dependent" analysis.

1. *The thesis*

Let me begin, somewhat oddly, at the end of Shoemaker's essay. Shoemaker devotes considerable space in his paper towards the development of the "Response-dependence" thesis in question, but hardly any space at all towards explaining how that thesis has implications for the substantive dispute that prompted Strawson's essay. Nevertheless, it is clear that Shoemaker thinks that this thesis *does* have such implications. Shoemaker concludes by saying:

> If I am right, then investigating the nature of the blameworthy (in the accountability domain) reduces to a matter of investigating the fittingness conditions of anger. Some theorists already engage in this general method, but often it is done only as a way of revealing what response-independent properties our emotional responses allegedly track. On their approach, it is an open question whether our set of responses might need or lack independent justification, and so an open question whether, for instance, determinism's truth could undermine responsibility. On my approach, asking for an

[1] Beglin 2018, De Mesel and Heyndels 2019, De Mesel 2022(a). Cf. also Balaguer 2023. There is, I believe, no one stable meaning of "the reversal" in the literature; sometimes this thesis appears to be a broadly methodological thesis, and sometimes it appears to be a kind of metaphysical thesis.

[2] For one recent survey, see Lopez de Sa 2013.

> external (response-independent) justification of the angerworthy would, as Strawson said, miss the point, for it would be to ask for a justification for being human. (2017: 521)

I lack the space to fully unpack this complex passage (although I will return to it below). But let me begin by noting Shoemaker's key contention. On one kind of approach, so the thought goes, it is an open question whether determinism is consistent with responsibility. On *Shoemaker's* approach, he contends, this question is closed.

But what approach? We can start with what is now a widely assumed biconditional linking *blameworthiness* to the appropriateness of *blame* – or what Shoemaker calls moral *anger*. (The difference, if any, between blame and moral anger won't matter for our discussion here, and I will use these terms interchangeably.)

> (T) S is blameworthy with respect to A iff S merits blame with respect to A/it is appropriate to blame S with respect to A/it is fitting that S should be blamed with respect to A.

(T) alone merely establishes the *extensional equivalence* of two concepts/properties. But we can ask a further question: which side of (T), if any, has priority? In other words, we can ask our Euthyphro question:

> <u>Euthyphro question</u>: Is S blameworthy because S merits blame, or does S merit blame because S is blameworthy?

The "traditional" order of explanation here is clear: the explanatory direction moves from a subject's being *blameworthy* to the appropriateness of *blaming* that subject. Now, various authors (writing under the "Strawsonian" banner) have suggested a *reversal* of this order of explanation: it is not that it is appropriate to blame S because S is blameworthy; rather, if S is blameworthy, this is because it is appropriate that S is blamed:

> (S) S is blameworthy with respect to A iff <u>and because</u> S merits blame with respect to A.

We can now make our first observation: (S) may be plausible[3], but it is completely silent on the *conditions* of merited blame, and so completely silent on the debate between compatibilists and incompatibilists (Todd 2016). The incompatibilist can plainly grant that someone is blameworthy if and only if *and because* it is appropriate that she be blamed – but then add that no one in a deterministic world is appropriately blamed.

Shoemaker's aim is thus to give our responses some key role beyond the role they play in (S). What role is that? As a first approximation, the idea is to give our responses a role in determining *what the conditions are* for appropriate (or fitting) blame.[4] In other words, we need to ask some key questions at a different *explanatory level* than that at issue in (S). And here we come to the idea of "response-dependence". We can start with a *non-normative* response-dependent thesis – the thesis Shoemaker labels "Dispositional Response-Dependence" (and here I quote):

> (DRD) X is blameworthy (and thus responsible) for some action or attitude A if and only if, and in virtue of the fact that, people are disposed to respond to X with blaming anger for A in certain standard conditions.[5] (2022: 315)

On (DRD), the facts about the conditions of blameworthiness ultimately reduce to merely *descriptive facts* about what certain actual human responses would be under appropriate conditions. The intent behind (DRD) is fairly clear, but we need to ask two key questions that (in some guise) will recur at various points in this essay. First, as stated, (DRD) – rather enigmatically – simply refers to "people"; if *people* are disposed to respond to X with blaming anger (in the relevant conditions), then X is blameworthy. Well, which "people" are relevant? Second, (DRD) – like any response-dependent analysis of any concept/property – appeals to

[3] For a defense of (S), see Menges 2017.

[4] The intent seems to be to give a "criterion thesis" in the sense of De Mesel and Heyndels 2019.

[5] On Shoemaker's presentation, this view is parallel to a similar view about *the funny*, viz:

> X is funny if and only if, and in virtue of the fact that, people are disposed
> to respond to X with amusement in standard conditions. (2022: 311)

These are the relevant "basic equations", in the terminology of Johnston 1989, where the term "response-dependence" was first introduced.

"standard" or "appropriate" conditions. (I will use these terms interchangeably.) Which conditions are appropriate in this sense? We don't need an exhaustive specification – which in any case is impossible to provide – but we do need a rough idea.

Let's start with the first question. Plainly enough, "people" in this analysis can mean neither *some people* nor *all people*; the former is clearly too weak, and the latter clearly too strong. Surely it isn't enough to render S blameworthy that *some few people* are disposed to respond to S with anger (in the relevant conditions), although everyone else isn't, and surely it isn't enough to *absolve* S from blameworthiness that some few people would *not* respond to S with anger in those conditions, although everyone else would. In this analysis, then, "people" can mean neither (literally) some people nor (literally) all people. What it must mean – I suggest – is thus something like *statistically normal people.* There are, of course, highly non-trivial questions about how to obtain the relevant class of "statistically normal people", but suppose we waive this concern. A more serious difficulty arises when we observe that the statistically normal people could *disagree*: once we've obtained our class of statistically normal people, it could turn out that 30 percent have the relevant disposition to degree 1, 30 percent have it to degree 0, and the final 40 percent have it degree .5. (That is, they are *somewhat* disposed to anger – or perhaps they are fully disposed to have a lesser *degree* of anger.) A thesis in the style of (DRD) needs to give us a verdict here.[6] What is it? Do we somehow average these responses themselves? Perhaps this is a promising (albeit in many ways highly artificial) way forward, but strictly speaking this is to abandon (DRD); whether S is blameworthy isn't fixed by whether statistically normal people are disposed to be angry with S (in the relevant conditions), but instead by the degree to which the *statistically average response* holds her responsible (in those conditions). There are difficult questions here.

Second, what are "appropriate conditions"? Plainly, the idea behind (DRD) is not that whether someone is blameworthy is determined by whether normal people would be inclined to blame her *while depressed*, or *while drunk*, or *while under threat*, or *while worried that unless this*

[6] The problem here is certainly an ancient one; cf. Socrates' challenge to Euthyphro's suggestion that the pious is *what the gods are disposed to love*: what if the gods disagree? (7e) One natural thought: perhaps *is blameworthy* is simply vague, and there are borderline cases. Here I must set this issue aside. Of course, one problem concerns the possibility that the statistically normal people could disagree, but one might have worries about the normative importance of statistical normalcy in the first place; cf. Fricke 2015. Note: the role of statistics looms large in Hieronymi's 2020 defense of what she considers to be Strawson's key point that "abnormality cannot be the universal condition" – but the role of statistics in *this* argument (as presented by Hieronymi) is very different than the role of statistics as discussed here (cf. Darwall 2021 and Russell 2021 for further discussion). But I must set these issues aside.

person is held responsible there will be a dreadful calamity, or *while mistakenly thinking that this is the person that committed the crime.* The appropriate conditions are thus at least to some extent "idealized". So far so good. Notably then, the appropriate conditions thus involve – to some first approximation – full empirical knowledge in conditions in which nothing "forward-looking" is known to be at stake. An example may help. Consider the tragic case of Robert Harris.[7] Now, it seems plausible that the vast majority of normal adult human beings would be disposed to be angry with Harris *directly* on seeing Harris commit his crimes *on the day he committed them, while knowing nothing further.* If these conditions are the relevant ones, then a suitably refined thesis in the style of (DRD) will (very plausibly) tell us that Harris is blameworthy. But are these the relevant conditions? Arguably not, for the obvious reason that many (most?) adult human beings would (and do) become much more ambivalent about Harris' responsibility upon becoming apprised of his childhood trauma. But surely it is our dispositions to anger in the *latter* circumstances that matter – that is, our dispositions once all the empirical facts of the case are in and duly considered. One's blameworthiness cannot be determined by a jury that doesn't even know the full facts of the case. (This realization will become important below.)

Let's now return to (DRD). Shoemaker raises several problems for (DRD), the chief problem being that its *analysans* lacks *normativity.* There are several issues here. First, it is perhaps directly implausible that the facts about the conditions of blameworthiness should reduce to non-normative, merely descriptive facts about the patterns of response of actual human persons. Second, there are problems of *variation*: if the facts reduce in this way, then it seems to follow that had statistically normal adults been disposed to blame young children (or the seriously mentally ill), then there would have been no condition on blameworthiness which prevents young children from being blameworthy. And that can seem like the wrong result.

There are perhaps ways of addressing this latter worry[8], but Shoemaker ultimately counsels abandoning (DRD), and instead suggests that we endorse a *normative* "response-dependence" thesis as follows:

[7] As famously discussed in Watson 1987.

[8] E.g., by "rigidifying" (cf. Vallentyne 1996) so that our actual responses fix the standards in nearby worlds as well; in that case, we might contend that the relevant counterfactual ("Had [contrary to fact] we been disposed to blame children, they'd be blameworthy") is *false*. An obvious reply: the relevant theory is still committed to the truth of the *indicative*, "If people are in fact disposed to blame children, then children are blameworthy." And this again seems objectionable. (Compare: even if, on similar grounds, the divine command theorist isn't committed to "Had God commanded murder, murder

> (NDRD) X is blameworthy (and thus responsible) for some action or attitude A if and only if, and in virtue of the fact that, *the person with the refined anger sensibility* would be disposed to respond to X with blaming anger for A in certain standard conditions.

As Shoemaker explains,

> This means that, when the person with the refined sensibility has a clear-eyed view of the matter, isn't tired or depressed, and isn't under some other distorting influence, her responses are simply what constitute the fitmakers, and so *determine the relevant oughts*, the reasons we have to be amused (in the case of the funny) or to be angry (in the case of blaming anger).[9] (2022: 319)

Thus, instead of appealing to "people" – or statistically normal people – we appeal to "*the person* with the refined sensibility"? But just as we asked about who the "people" are relevant to (DRD), we can ask who the *person* is relevant to (NDRD). Plainly, of course, the thought behind (NDRD) isn't to appeal to any particular concrete person at all. Since (NDRD) is meant to be primitively normative, the idea is that we simply posit, as normatively basic, the uniquely refined anger-sensibility; the standards of blameworthiness are thus fixed according to whether this sensibility is triggered (or not) in the relevant conditions.[10]

would be right," she nevertheless *is* committed to, "If God in fact commands murder, murder is right" – which again seems unacceptable.) Sidenote: the relevant phrase, "had statistically normal adults been disposed to blame…" contains an important *de re/de dicto* ambiguity. If we read this claim *de re*, we fix on some concrete individuals, say A, B, and C, who are the statistically normal adults, and the claim is then that if A, B, and C had been disposed to blame children, children would have been blameworthy. Read *de dicto*, however, we are just saying that had "statistically normal adults" referred to some different class of adults, such that *that* class of adults had the given disposition, then children would be blameworthy. It is the *de dicto* reading that seems relevant here.

[9] Shoemaker surprisingly never explicitly formulates (NDRD) – unlike the earlier (DRD) (which was a direct quote). This may be due to an issue also noted by Heyndels and De Mesel (2018). Oddly, Shoemaker says that a chief problem for (DRD) is that it is "completely unclear" what the "appropriate conditions" are relevant to this thesis. But even if this were a deep problem for (DRD), Shoemaker seems not to notice that the exact same questions will arise concerning his preferred *normative* response-dependence thesis; the fact that the thesis is *normative* does not obviate the need to appeal to "appropriate conditions". Note, after, all, Shoemaker's formulation: "isn't tired or depressed, and isn't under some other distorting influence…". This is precisely the appeal to "appropriate conditions" – and formulating the thesis explicitly (as in (NDRD) would have made the need for this appeal salient.

[10] One key question – which I will not here address – is whether a *normative* response-dependence thesis respects the dialectical motivations of attempting to go "response-dependent" in the first place. If one's

But now we can make a simple observation. Unless supplemented in some way by further considerations, (NDRD) appears to have no compatibilist implications whatsoever. Indeed, it seems obvious that an incompatibilist could simply accept (NDRD), and go on to contend that no one with a truly *refined* anger-sensibility would (in appropriate conditions) be angry with someone whose actions he or she knew to be determined. In point of fact, D'Arms and Jacobsen – whose discussion Shoemaker seems to follow (and often cites) – offers it as an *advantage* of a normative response-dependent theory that it can explain exactly the sorts of disagreements we see between incompatibilists and compatibilists:

> The focus on merited responses also promises to explain what is at issue in the many seemingly cogent evaluative disputes that cannot settled by empirical investigation: these are disagreements over what response is merited. Disputants can criticize as unmerited even very common patterns of response — such as survivor guilt, fear of spiders, and regret over the bad outcomes of good decisions — and they can back up these claims with reasons that will sometimes be persuasive. (2006: 203)

The incompatibilist can criticize as unmerited what could even be a very common pattern of response – viz., the tendency to blame those whose actions were determined – and they can back up these claims with reasons that appear at least to some to be persuasive. The lesson is clear. The normative response-dependence thesis can perhaps help to *frame* the debate between the incompatibilist and the compatibilists, but it certainly cannot *settle* that debate – and *prima facie*, it cannot even *contribute* to that debate, on the plausible contention that incompatibilist argumentation can be construed as implicit argumentation to the effect that no truly *refined* person would be angry (in appropriate conditions!) with someone whom he or she knows was causally determined.

2. *An analogy and an argument*

motivations are in part to avoid the ghostly spectre of brute, Platonic facts about the genuine conditions of blameworthiness, it appears to be little improvement to posit a brute, Platonic fact about which sensibility is the "refined" one. Cf. Enoch 2005 on precisely this dialectical tension for certain "idealized" response-dependent theories in the normative domain.

As stated, then, the (NDRD) thesis has no compatibilist/non-skeptical implications. In much of Shoemaker's discussion, however, one can discern a sort of *narrow* construal of "response-dependence", and a broad construal. The narrow construal is one that simply consists in the truth of (DRD) or (NDRD). However, at times, Shoemaker links "response-dependence" to some larger (and often opaque) Strawsonian themes, especially the theme that it is some kind of mistake to ask for an "external justification" of our practices of responsibility, or to provide an "external criticism" of those practices. (The key difference between *those* notions is itself in need of elucidation, but more on this to come.) Now, it may be that, if we somehow identify the response-dependence thesis with the thesis that it is a mistake to provide this kind of "external criticism" – or show that there is an implication from the former thesis to the latter – then we may conceivably get a compatibilist upshot, if the incompatibilist can be seen to be providing an "external criticism" in this sense. (Something like this suggestion appears latent in the passage initially quoted above.[11])

So let's back up. I suggest we revisit an analogy suggested by Todd (2016) between *the blameworthy* and *the funny* – an analogy that itself structures and motivates much of Shoemaker's discussion. Now, the key observation Todd makes is simple: very plausibly, it couldn't turn out that nothing we've ever been laughing at – more carefully, been comically amused by – has really been funny. If the blameworthy is relevantly similar to the funny, then perhaps it similarly couldn't turn out that no one we have ever blamed has really been blameworthy. This strategy appears both simple and potentially decisive. In what follows, I aim to show that it is neither.

But first a detour. The notion of "response-dependence" has been associated with a certain sort of phenomena, one involving a dynamic between the possibility of *individual* error but the impossibility of *global* error. Again, consider *the funny*, or (in other words) the *amusement-worthy*. Now, it is worth observing that – plausibly – certain individuals can be mistaken to think that some token item is (or isn't) funny. As Crispin Wright has noted (in connection with a response-dependent theory of the funny), there is nothing funny about what

[11] Unfortunately, I lack the space to do this passage sufficient justice – but in my estimation, it presents us with several interpretive difficulties, of which I mention just two. Shoemaker, recall, says that "asking for an external (response-independent) justification of the angerworthy would, as Strawson said, miss the point." But (1) it isn't obvious that "justifications" can be "response-(in)dependent" (it is *concepts* or *properties* that can be response-(in)dependent), and (2) it isn't clear what it could be to offer a justification of "the angerworthy"; one cannot justify a property.

happened at Chernobyl, even if some (unfortunate) individuals *think* this is funny. (2003: 32) Nevertheless, there is a certain sense in which it *isn't* possible that *everyone* should always have been mistaken in this way – that is, that nothing *anyone* has laughed at has been funny. Or consider McGeer's example of the *fashionable* (in the domain of clothing): Jack may be mistaken to think that his wardrobe is fashionable. But somehow it *isn't* possible that no wardrobe has *ever* been fashionable. (2019: 303) Well, what exactly is going on here? As a first approximation, we want to be "realists" enough about the relevant properties to allow for errors, but ... anti-realists? Or "constructivists"? Or ...? ... about the relevant properties to eliminate the possibility of *global error.* Can we thread this needle? Perhaps we can.[12]

My sense is that drawing out this comparison is the best hope that a (broadly) "response-dependent" theory can have anti-skeptical implications. To help us structure the discussion, let us put the argument schematically as follows:

1. The standards of blameworthiness are determined in the same way as the standards for funniness.
2. It couldn't turn out that nothing we've laughed at has been funny.
3. If (1) and (2), it couldn't turn out that no one we've blamed has been blameworthy.
4. It couldn't turn out that no one we've blamed has been blameworthy.

In this paper, I want to grant as much as possible to the thought behind (1). And of course I want to grant the obvious thought behind (2). The trouble, I contend, is (3). The funny and the blameworthy can be (in this key way) normatively similar, but differ along a further important dimension. And this key difference undermines the move from (1) and (2) to (4).[13] More to the point: it isn't *simply* because the standards of funniness are determined in the way

[12] "Response-dependence" has also been associated with *individual* infallibility (in appropriate conditions); indeed, Pettit (1991: 622) writes that, "[A]s an observer under normal ["appropriate"] conditions cannot be in ignorance or error about the colour of something, so the responses involved in any response-dependent area of discourse cannot lead subjects astray under those conditions." Cf. Wright 1989 and Holton 1991 on a response-dependent account of *intention*, and also Holton 1992. Here I must set these connections aside.

[13] Note: the conclusion of this argument is *non-skepticism.* There is, of course, a non-trivial gap between non-skepticism and compatibilism; in order to close it, one would have to have an argument against *libertarianism.* But what is that? The Strawsonian might suggest several such arguments, e.g. an *epistemic* argument to the effect that libertarianism goes beyond the "facts as we know them". For a recent assessment of this style of argument, see Todd and Rabern 2023. But I set this aside.

that they are that (2) is true; it is because of this fact, and a key *further* fact, viz., that *the funny* has no non-trivial empirical presuppositions. And this marks a key difference between the funny and the blameworthy.[14]

3. *Four examples*

To see this result, we have to see that there are at least *two* possible sources of global error with respect to some concept/property:

> (A) It could be that our standards for application of that concept/property are mistaken
>
> (B) It could be that even if our standards for application of that concept/property are *not* mistaken, we are simply factually mistaken to think that anything has ever met those standards.

My key claim is simple. What the "impossibility of global error" thesis allows us to eliminate – on *a priori* grounds – is possibility (A). But it leaves open possibility (B). And therein lies the trouble. An analogy might help. Imagine we've failed to reach the destination we were trying to reach. There are (at least) two possibilities. The first is that our map was inaccurate. The second, however, is that even though our map was accurate, we made factual errors in following it. What the relevant thesis allows us to dismiss is the idea that our map, all along, has been an inaccurate map of the given terrain – perhaps, mysteriously, because somehow our map *determines* the terrain. But that thesis would leave open the second possibility, the possibility that we are bad at telling when we've followed correct directions.[15]

[14] We might consider a nearby argument that proceeds, not *via* an analogy with the funny, but instead via the analogy with *the fashionable*. Indeed, there might be a principled Strawsonian reason to pursue this latter comparison rather than the former. Strawsonians often emphasize the role of our blaming *practices* – the key idea (perhaps) being that there cannot be an external criticism of the standards operative in this *social practice*. But note that humour isn't a "social practice" in any obvious sense – whereas of course fashion is clearly a social practice. At any rate, my sense – which I cannot here justify – is that, despite this difference, a closely parallel dialectic would unfold vis-à-vis *the fashionable* as we will see shortly vis-à-vis *the funny*.

[15] De Mesel (2022: 1907 - 10) maintains that what he calls "Strawson's view" allows for certain forms of collective mistakes about responsibility. First, we can be mistaken about what the rules of the practice really are. Second, we can be (factually) mistaken about whether the rules apply in particular cases. Surprisingly, however, De Mesel doesn't go on to apply this second point to Strawson's purported chief concern in "Freedom and Resentment", viz., the debate between the compatibilist and the incompatibilist. That is, in effect, my aim in what follows.

We need to consider some examples that will allow us to observe certain key patterns. Here I will employ the following heuristic. Imagine that aliens appear, and they tell us that they have been carefully observing us over (say) the last 100 years, and have accordingly learned our language, which we verify to our satisfaction. (A heuristic is needed, and I haven't found a more realistic one than this.) Now, first, imagine the aliens saying the following:

> 1. *We have inspected your new spaceships, and none of these spaceships are interstellar-travel-worthy.* [Imagine we've been building spaceships intended for interstellar travel (IST)].

If we are highly confident in our (as-yet-untested) spaceships, we of course may be *skeptical* of what these aliens say. But what they say is perfectly intelligible. And note that what they say could be true for two different reasons, working either individually or in combination. First, the aliens could tell us that our standards for what counts as IST-worthy are simply the wrong standards; we've been thinking that if an object meets conditions *C1* – *Cn*, that will render that object IST-worthy. But we're totally wrong about this. At any rate, we certainly cannot rule this out *a priori.* (Note: it would be absurd to try to say, "But the standards for what counts as IST-worthy are simply fixed by our dispositions to regard things as worthy of IST!" The IST-worthy is in no sense "response-dependent.") Second, the aliens could tell us that we've been right to think that if an object meets conditions *C1* – *Cn*, that will render it IST-worthy. But they tell us that we have been very bad at telling when objects meet these conditions: contrary to our beliefs, none of our spaceships do.

Thus, return to our two sources of global error. And now we can ask (and answer):

A. Could our standards be wrong? (In principle, yes. We cannot rule this out *a priori*.)
B. Could we be wrong to think that anything has met those standards? (In principle, yes. We cannot rule this out *a posteriori.*)

Global error about the IST-worthy – about what is fitting to be used for IST – is thus possible twice over. Our standards could be wrong, and since the conditions that render an item IST-worthy have non-trivial empirical presuppositions, we could also be factually mistaken to think that *any* token item has met those standards. But let us now contrast *that* case with a key case at issue in this paper, viz., *the funny*. Suppose the aliens come and say:

> 2. *We have listened to your comics and watched your movies, and more else besides. But nothing you all have ever laughed at (been amused by) has really been funny* (*amusement-worthy*).

Just as the above is wholly intelligible, this is instead wholly confusing. (One may say that this is *itself* funny.) First, suppose the aliens try to say that our standards for being amused are simply the "wrong" standards. But what could this even mean? It can't be that our standards are the "wrong" standards. If the aliens say that our standards are the wrong standards, we can reply that they are simple changing the subject. It is perhaps a difficult question *why* our standards can't be wrong – indeed, perhaps this result is overdetermined – but the key point at this stage is simply to grant that they *can't* be wrong.[16]

Now, could it be that we have always and everywhere been factually mistaken to think that anything meets those standards? In the case of *the funny*, the answer, plausibly, is "no". Suppose the aliens grant that our "sense of humor" (so to speak) is perfectly fine; they just claim that there has always been something about what we've been laughing at according to which, *given* our sense of humor (that is, by our own standards) that thing actually isn't funny.

[16] Note: here and elsewhere, we need to distinguish between our *theory* of the standards that operate in our practice, and the actual truth about the standards that operate in our practice. The idea here is that the relevant standards themselves can't be wrong; it isn't that our *theory* of these standards can't be wrong. (Compare: "your theory of the rules of chess is wrong" vs. "the rules of your game of chess are wrong".) For discussion, see De Mesel 2022: 1907. A further note: talk of "standards" in the case of *the funny* is perhaps slightly artificial. More to the point, it isn't even clear that (we think that) there *are* any objective conditions C*1* – C*n* such that, if something meets those conditions, that will render it funny. (Shoemaker himself [2017] makes a good case for something at least very much like this contention.) Nevertheless, talk of "standards" here, I assume, cannot be wholly out of place, given that we want to preserve individual fallibility: to pick up on Wright's example, an unadorned, mere description of certain innocent people suffering from radiation sickness *cannot* be funny, in which case there is some "condition" on funniness that such a mere description doesn't meet, something like *not being an unadorned description of horrible suffering,* or whatever. Note: this is not to say that it would be impossible for a comic to make a funny joke *about* radiation sickness. But there is a difference between Jack's radiation sickness *in itself,* and what could be a funny joke *about* such sickness. At any rate, my assumption here is that it is sometimes strictly *true* to say, "That simply wasn't funny," even if some individual *found* it funny. If this claim is itself false, then premise (1) in our key argument is a non-starter, for there is – by everyone's lights – an obvious (token-level) difference between being *found* blameworthy, and being blameworthy. More to the point, if this is false, then the reason we could dismiss the alien's suggestion is simply that there is no space between *being found funny* and *being funny*, in which case the only way they could show that nothing has ever been funny is by showing that no one has ever found anything funny – which, of course, they can't do, leaving in place "the facts as we know them". It is simply an obvious datum that sometimes some of us *have* found some things funny. It is similarly an obvious datum that sometimes some of us have *found* some people blameworthy (i.e., blamed them) – but this alone, I am assuming, does not settle the question of skepticism.

But how would this go? It is, of course, completely unclear. But by way of illustration, suppose they say: actually, you didn't know it, but all of these jokes were told by deplorable racists. Well, maybe that implies that we shouldn't have laughed at those jokes – but it doesn't imply that the jokes weren't funny (this would be the "moralistic fallacy"[17]). Or imagine that they reveal that the pun in your friend's email that you just laughed at was no such pun at all – indeed, the text was hammered out by monkeys on typewriters. Well, *something* was still funny here, even if it wasn't precisely what you had thought it was. The *appearance as of there being this pun* was funny, even if there was no pun. Thus, it is completely unclear what the aliens could reveal that would show that – by our own standards – nothing we've ever laughed at has been funny.[18]

We thus have the inversion of the case considered above:

A. Could our standards be wrong? (No, not even in principle. We can rule that out *a priori.*)
B. Could we be wrong to think that anything has met those standards? (No. We can rule that out *a posteriori.*)

Hence, global error about the funny is ruled out. But now consider a further case, a case which plausibly bears important resemblance to the interstellar-travel case, *and* the funniness case. For simplicity, imagine that the relevant community is a religious one, and consider:

3. *Nothing you all have worshipped has ever really been worship-worthy.*

[17] D'Arms and Jacobsen 2000.

[18] Recall the thought experiment: the aliens have been observing us *as we have been* over the last 100 years; it isn't relevant here to observe that the aliens could (in some sense of "could") reveal that the world was created 5 minutes ago. That is to raise the spectre of garden-variety epistemic skepticism, not *responsibility* skepticism. Related: I am not claiming that it is metaphysically impossible that there should be a world in which there exist only persons S1 – Sn, these persons are comically amused by various token items, but none of these items are funny. Perhaps there is a world in which there is one and only one episode of amusement, and that one token episode is one for which it is right to say, "That wasn't funny, even though you found it funny". In that case, it turns out that though there had been amusement in this world, nothing anyone was amused by in this world was amusement-worthy. I am instead (i) holding fixed the "facts as we know them", and claiming that (ii) there is no way of providing any kind of "back story" – however elaborate – which would *reveal* that nothing we've been laughing at has really been funny.

This is *potentially* confusing, but not necessarily confusing. If they allege that our standards of worship have been the "wrong" standards, then perhaps we can insist that whether something counts as worship-worthy is simply fixed by our dispositions – say – to fully devote our lives to it. At any rate, I shall grant that, for one reason or another, there cannot be any kind of "external criticism" of our standards of worship. But unlike in the case of the funny, it is not confusing at all if they allege that, even if our standards of worship-worthiness are perfectly fine, we have been factually mistaken to think that anything has ever met those standards. Maybe they say, "We're saying that, contrary to what you've been thinking, none of the objects you have been worshipping actually meet the standards you have in mind for worship-worthiness; in point of fact, these objects [e.g., gods, God] have been sophisticated holograms of our own making, or they simply fail to exist in the first place." In this case, they aren't criticizing our standards; they are saying that nothing meets them, contrary to our beliefs. (Are they mounting an "external criticism" of our "practices of worship"? That's ambiguous; more on this below.) Standard atheists may be entirely happy to grant that *if* there existed a being that met the description at issue in theistic religion, that being would be worthy of worship. The mistake of the religious, therefore, is not (or not necessarily) having the wrong standards of worship; their mistake is instead straightforwardly factual. Thus again our two questions:

> A. Could our standards be wrong? (No. We can rule that out *a priori.*)
>
> B. Could we be wrong to think that anything has met those standards? (Yes! We *cannot* rule that out *a posteriori.*)

Hence, global error about the worship-worthy is *not* ruled out. (Indeed, I suspect that most readers of this paper are prepared to say that not only is it not ruled out, but that it is also *actual.*) Like the funny, it couldn't be that our standards are the wrong standards – but like the IST-worthy, being worship-worthy, *according to our own standards*, requires meeting substantial non-trivial empirical conditions, and in principle we could be globally mistaken to think that anything has ever met those conditions.

Our question now comes into view. Is the blameworthy more like *the amusement-worthy*, or instead more like *the worship-worthy*? That is, we can now ask our central, crucial question. Imagine the aliens say the following.

> 4. *No one you have ever blamed has been blameworthy.*

Now, let us simply grant that we can reply, as in the previous two cases, that it can't be that our standards for what counts as blameworthy are the "wrong" standards. There can't be this kind of "external criticism" of our standards (perhaps because our dispositions to blame somehow *determine* the standards, or for some other reason). Now the question becomes: is blameworthiness more like *being funny* (no non-trivial empirical presuppositions), or instead more like worship-worthiness (substantial empirical presuppositions)? In other words, we can't be making a *normative* mistake concerning our standards – but could we be making a *factual* mistake to think they've ever been met? Well, suppose the aliens said the following:

> As we've said, we have been observing you carefully for years. And by your own standards – that is, according to your own practice – people aren't blameworthy if what they've done is unavoidable for them. But we can now reveal that everything has always been determined, and so nothing has ever been avoidable. So it turns out that no one has ever met the standards of blameworthiness *according to your own practice.* Thus, no one you have ever blamed has really been blameworthy.

What can we say back? More carefully, what can we say back if we want to show that what these aliens say is *false*? I can see three and only three options. Unsurprisingly, they are entirely familiar:

> (a) you are wrong about us; we aren't determined. (Libertarianism)
> (b) you are wrong about the practice; avoidability is no part of the practice (the rejection of the Principle of Alternate Possibilities [PAP])
> (c) you are wrong about determinism; it doesn't imply that nothing has ever been avoidable (classical compatibilism).

But now the key point. Prima facie, we are now simply *back to the same bitter disputes.*[19] Do our practices encode an "avoidability" condition on responsibility (i.e., PAP)? Some say no, but

[19] Note: the point of the aliens' speech here is to articulate a standard global challenge from the truth of determinism to the claim that no one is blameworthy. I have put this challenge in terms of *avoidability*,

there is certainly room for debate.[20] And if there is such a condition, does determinism preclude avoidability? Again, some say no, but the question is far from settled.[21] By all appearances, then, the relevant Strawsonian understanding – whatever precisely that comes to – has gotten us nowhere. And this is true even if we concede the key insight at stake, which the "response-dependence" thesis is meant to secure, viz., that it cannot be that the standards implicit in our practices are the wrong standards. Could our standards be wrong? No. But could it be that no one has ever met them? That possibility hasn't yet been ruled out – and ruling it out seemingly requires re-litigating the familiar debates.[22]

but we could also put it in terms of "sourcehood" (see, e.g., Pereboom 2014: Ch. 1) – and the dialectic would unfold similarly: either the sourcehood condition isn't part of the practice, or instead determinism doesn't preclude sourcehood.

[20] For "no": Wallace 1994: Chs. 5 and 6, and Fischer and Ravizza 1998; for a recent "yes" on the compatibilist side: Brink 2021. (And for a "yes" on the incompatibilist side, see Todd 2017.

[21] In a recent article, De Mesel (2022(b)) convincingly shows that P.F. Strawson's own position is (c), i.e. classical compatibilism. It is, of course, beyond the scope of this paper to evaluate classical compatibilism, but De Mesel's discussion is in line with my own chief conclusions: ultimately, even by Strawsonian lights, we cannot *avoid* the chief (sometimes pejoratively called "metaphysical") debates we see in the compatibility literature. Those debates must be faced head-on.

[22] Fischer 2014 argues (correctly to my mind) that Strawson does not succeed in fully "sequestering" the relevance of "metaphysics" to the question of whether anyone is morally responsible. Ultimately, Fischer suggests that Strawson's position is vulnerable to traditional incompatibilist worries about "could have done otherwise" and manipulation – or at any rate that Strawson at the least does not add to our understanding of how these traditional worries can be addressed. Fischer writes:

> Insofar as Strawson holds that our reactive attitudes are triggered simply by the quality of others' will, quite apart from questions about the history behind the display of the relevant attitudes by others, then his approach is singularly unable to address yet another important and central worry of the incompatibilist: the challenge of manipulation. (111)

And with this I agree (cf. Todd 2011, 2013). Fischer adds:

> In contrast, to the extent that my approach is a historical approach that embodies the requirement that the actual-sequence mechanism issuing in the behavior in question be *the agent's own mechanism*, it can (arguably, at least) address the incompatibilist's concern. (111)

Arguably – but arguably not. It seems to me that Fischer is right about Strawson, but that his own position is nevertheless dialectically unstable. If the problem for Strawson's view is that poor quality of will can simply be "manipulated in", then it seems to me to be mysterious how we might solve the problem by imposing yet *further* conditions that can *also* be "manipulated in" – like Fischer's own "taking responsibility" condition. Of course, it is not strictly *incoherent* to maintain that once someone is manipulated into taking responsibility for the mechanism that issues in her poor quality of will, *then* she is responsible. But this position nevertheless strikes me as vaguely unprincipled. These issues are delicate, however, and in fairness, Fischer and Ravizza are certainly aware of this concern and seek to address it (1998: 236 - 7). The worry about manipulation, in my (admittedly unsupported) judgment, ultimately points us towards the necessity of some condition that by its very nature cannot be

4. *A compatibilist way forward?*

But let’s slow down. My contention is that in response to the relevant challenge, the compatibilist must say either that avoidability is no part of the practice, or instead contend that determinism is no threat to avoidability. But suppose we consolidate these two responses into one. On the one hand, the compatibilist can concede that there is *some sense* of avoidability that is part of our practice, but she will contend that *this* sense of avoidability is consistent with determinism, even if some other kind of avoidability – what we might call *all-in* avoidability – is not. This posture amounts to the position that a condition of *all-in* avoidability is no part of our practice. And now the compatibilist may feel like she sees an opening – something that is problematically *off* about the alien’s speech above. How have the aliens come to know about our practice what they evidently purport to know – viz., that our practice encodes a condition of *all-in* avoidability? Indeed, we can observe that, if the alien’s speech is really to trouble us, it must be recast as follows:

> As we’ve said, we have been observing you carefully for years. And by your own standards – that is, according to your own practice – people aren’t blameworthy if what they’ve done is *all-in* unavoidable for them. But we can now reveal that everything has always been determined, and so nothing has ever been *all-in* avoidable.

But now the compatibilist might – all in a rush – interject that we never in fact excuse on grounds that what someone did was unavoidable in *this* sense. How then could their *observations* support the view that our practices involve a (perhaps implicit) condition of all-in avoidability? Indeed, compatibilists are often at pains to emphasize exactly this point: we never excuse on grounds that what the relevant agent did was determined (all-in unavoidable). But rarely has there been a more influential observation in the responsibility literature that is at the same time so obviously facile. Yes, we never in fact do so, but that is perhaps because we do not in general *believe* that what people do is unavoidable in this sense. It is, accordingly, highly non-obvious how we would react (in the relevant conditions), given a convincing

manipulated in from the outside – and that condition, if it can be formulated coherently at all, will be libertarian.

demonstration of the *truth* of this belief. More concretely, suppose our aliens, now apologizing for getting ahead themselves, added something like the following:

> We have abducted a statistically significant random sampling of normal adult human beings and put them into a device similar to what you have called "the experience machine". Once plugged into the experience machine, we carefully tested certain hypotheses, the chief of which being whether these human beings would be disposed to anger with certain individuals (in certain key conditions) upon being given – to their minds – a fully convincing demonstration that what the relevant individuals did (and what everyone else did as well) was causally determined "from the start". We noted the following results. When our sample of humans was in a calm mood, and they didn't feel like anything "forward-looking" was at stake, almost none of them were morally angry with the relevant individuals whom they now believed had been causally determined to be as they were and do what they did. Their attitude was instead something like what you may call ambivalent resignation; they were shocked and upset that the universe was so strange (and seemingly "unfair") – so that in order for anyone to have turned out any differently, the very facts at the Big Bang would have had to have been different. But once they processed their thought that this is how the universe really is, none of them were really disposed to anger.

Well, is this result the likely one, or is it not? Experimental philosophers may be thinking, "Grist for our mill! We can test this…", but this is no grist for the experimental mill. There is all the difference between being asked to *consider* a deterministic universe, or to *suppose* that our universe is deterministic, and to be given (what appears to be) a convincing *demonstration* that it *is* deterministic – and this is something no present-day experimental philosopher can accomplish. It is thus fundamentally unclear whether a condition of all-in avoidability is "part of our practice" in the relevant sense. Incompatibilists can – and have – put forward thought experiments designed to elicit the key judgments, but compatibilists can of course dig in their heels.[23] We have no easy way of settling this empirical question. But let me note one key

[23] In various places, Fischer insists that he would be completely unmoved, even if he were convinced by a consortium of physicists that determinism is true (e.g. Fischer 2023). I should perhaps let Fischer speak for Fischer, but it is important to see that our relevant dispositions after accepting a mere "announcement" of this sort isn't the key test. It is, after all, easy to compartmentalize, and a

confounder here. There is no reason why incompatibilist argumentation cannot be relevant to the very data our anthropologist aliens are collecting. In other words, suppose they said:

> In particular, we noted the following results. When our sample of humans was in a calm mood, and they didn't feel like anything "forward-looking" was at stake, *and had been exposed to certain philosophical arguments*, almost none of them were morally angry with the relevant individuals whom they now believed had been causally determined. That is, we observed that nearly all of our humans were initially confused about what to think or feel, but they then turned to philosophical reflection as a path forward. They reflected on what you have called the "consequence argument", the "basic argument", and the "manipulation argument"[24] – and various other arguments – and we then observed that almost none of them were disposed to anger.

What this would reveal is that it is *itself* part of the relevant "practice" to appeal to a standard of fairness (or the sense of all-in avoidability) that appears in incompatibilist argumentation. That is to say, when incompatibilists argue in their characteristic ways that it would be wrong or unfair or inappropriate to blame someone who is determined, they are ultimately not appealing to some standard of fairness that is external to the practice; indeed, quite the opposite – if what the aliens said here was true, it would turn out that it is itself *part* of the practice to appeal to exactly the kinds of incompatibilist arguments that some say are "external" to the practice. (Philosophizing is itself a human practice, we must remember.) But now our question is

disposition to assent to the proposition "determinism is true" needn't amount to the visceral knowledge *that determinism is true.* A comparison. Many ordinary persons assent to the proposition "there is a God", but there is an important difference between the ordinary contemporary believer and Moses after his (supposed) encounter at the burning bush; only the latter had (or is taken to have had) a fully visceral direct acquaintance with the truth of theism. It is one thing to remain angry with one's flatmate after processing a theoretical announcement "that determinism is true" – whatever that means! – but it is another to remain angry with him after processing a convincing demonstration, say, that what he did, just then, was the 10,000th rollback of a deterministic setup, in which – of course, and at the mere touch of a reset button – every time he does *exactly the same thing.* But I digress. Other esoteric examples might be considered; cf. the discovery of the universe-like "pods" described in Todd 2019. Of course, here we encounter some fraught dialectical issues that have arisen in connection with so-called "manipulation" arguments: are the esoteric examples parallel in the relevant ways to determinism? Here I must set these issues aside, although with one note: there ought to be *some* way of making determinism appropriately "visceral" – likely exploiting some element of *predictability* – and the question here concerns the dispositions of those who have integrated this visceral knowledge into their web of belief.

[24] Cf. van Inwagen 1983, G. Strawson 1994, and Pereboom 2014: Ch. 5, respectively.

complicated yet further: would unbiased people, in appropriate circumstances, be convinced by incompatibilist argumentation, or not? But of course this question cannot be settled in any obvious way: for those of us without access to an experience machine (and more else besides), it appears to be nothing further than the question of whether incompatibilist arguments are good arguments, and that is *exactly what is at stake*.

5. Judgments in Appropriate Conditions

Let us sum up the above discussion by approaching this issue from a slightly different angle. The intent behind a standard (non-normative) "response-dependent" analysis of a concept/property is to say the following: our reactions and/or judgments *in appropriate conditions* simply *determine* the application conditions for the concept (or the extension of the relevant property). In the "worship" case, this understanding gives us the following:

> There exists a being that is worthy of worship if and only if there exists some being such that, given appropriate conditions (e.g., full empirical knowledge, and…), normal adult human beings are disposed to worship that being.

Well, is there any such being? That is, of course, the question. It isn't clear. Theists will of course say yes; atheists/religious skeptics of course will say no. Similarly, the response-dependent understanding of responsibility gives us:

> There exists a person that is worthy of blame if and only if there exists some person such that, given appropriate conditions (e.g. full empirical knowledge, and…), normal adult human beings are disposed to blame (be morally angry with) that person.

Well, is there any such person, given that determinism is true? With full empirical knowledge, and nothing "forward-looking" at stake, are normal adults disposed to blame those they viscerally know to be determined? The answer to this question, like the one above, isn't obvious. Contrast these two cases with:

> There exists an item that is worthy of amusement if and only if there exists some item such that, given appropriate conditions (full empirical knowledge, and…), normal adult human beings are disposed to be amused by that item.

Is there any such item? The answer to this question, I assume, is an obvious "yes". In this light, consider one way we might put the key argument we have investigated:

> 1. The standards of blameworthiness operative in our practices of blame are encoded in our dispositions to blame in appropriate conditions.
> 2. These standards cannot be mistaken. Therefore,
> 3. There exists a person who is worthy of blame if and only if there exists a person we are disposed to blame in appropriate conditions.
> 4. There exists a person we are disposed to blame in appropriate conditions. So,
> 5. Someone has been blameworthy. (Non-skepticism)

The argument looks good at first pass. Premise (1) looks plausible – perhaps even trivially true. Premise (2) is, of course, highly controversial, but something I hereby wish to grant. (3) follows. The trouble, however, is (4). What tendency we might have to uncritically accept (4) plausibly stems from two sources. First, we are not carefully attentive to the fact that we are rarely (if ever!) in *appropriate conditions* in the relevant sense; for one thing, we do not possess full empirical knowledge, and if determinism is true, full empirical knowledge may affect our dispositions to blame. Second, there is our familiar problem, which proponents of this style of argument must somehow resolve: there is considerable ambiguity in how to understand the "our" in premise (1) and the "we" in premises (3) and (4); if we read this "we" as "some substantial number of us" (e.g., "we compatibilists"), then perhaps the claim is plausible – but that reading is too weak to support the weight of the argument.

It is worth quickly observing how a similar dialectic unfolds given a plausible development of the *normative* response-dependence thesis. Imagine our aliens coming down and saying, "We've listened to your comics, and none of them have a *refined* sense of humor," or more to the point, "We've observed you for some time, and none of you have *refined* anger-sensibilities." My sense is that a theorist like Shoemaker will be inclined to respond roughly as follows: it is a constraint on a plausible understanding of "refined" that some *actual human* *persons* are, if not *perfectly* refined, then at least very highly refined (whatever exactly that comes to). Given this thesis, we could reject the alien's suggestion – and to mount an argument for non-skepticism, we could thereby argue as follows:

1. Someone is blameworthy if and only if the person with a refined anger-sensibility would be angry with that person in appropriate conditions.
2. If most normal adults would be angry with S in appropriate conditions, then the person with a refined anger-sensibility would be angry with S in those conditions. (What is "refined" cannot come substantially apart from the dispositions of most normal actual human beings.) So,
3. There exists a person who is worthy of blame if and only if there exists a person who we – most normal adults – are disposed to blame in appropriate conditions.
4. There exists a person we are disposed to blame in appropriate conditions. So,
5. Someone has been blameworthy. (Non-skepticism)

Presumably everyone will grant (1), and many will reject (2). But even if (2) is granted, the trouble once more is (4).

6. *Interlude: the significance of framing*

My earlier contention was that, on reflection, the Strawsonian paradigm is getting us *nowhere*, and simply leaves in place all of the key debates that paradigm allegedly bypassed. That paradigm can *frame* the debate, but it cannot *settle* the debate. But perhaps we would be wrong to underappreciate the significance of exactly this framing. After all, observe that the key questions discussed above are (roughly) *empirical*: having granted premise (2) – that our standards can't be the wrong standards – the key question thus becomes the empirical one of whether *in fact* we would be disposed to blame in the relevant conditions. I have argued that it is highly non-obvious that these results would go compatibilist, and that the question of whether they *would* go compatibilist cannot easily be separated from the disputed prior question of whether unbiased people would be moved by incompatibilist arguments. But it is worth seeing the marked contrast between *this* debate, and a debate that might be prompted by a different speech our aliens could make, viz.:

> As we've said, we have been observing you carefully for years. And by your own standards – that is, according to your own practice – people are still blameworthy for what they've done, even if what they've done is all-in unavoidable for them. (We determined this *via* extensive testing in the experience machine.) But what this shows is that your practices are thus fundamentally unfair: people do not in fact deserve these responses in these conditions, contrary to the standards of your practice.

I must confess that I have a great deal of sympathy for what these aliens are saying here, conditional on the truth – which I doubt – of what they say concerning us and our dispositions. If we *are* in fact disposed in this way, then we are simply *wrong* to be disposed in this way (in the sense that we would be disposed to an incorrect response). Further, it isn't totally obvious why we should be entitled to dismiss this possibility – that our standards are the wrong ones, and that we – taken together – are sometimes disposed to reactive blame and anger when those attitudes are not in fact deserved. Accordingly, it is a highly significant result if we *can* in fact dismiss this possibility, precisely on grounds that this kind of "external criticism" of our standards is in principle illegitimate.

Conclusion

It is worth seeing where we've been in this essay. I have argued that a so-called "response-dependent" theory of responsibility does not deliver any compatibilist/non-skeptical conclusions. This much is, I suggest, obvious for a *normative* response-dependence thesis, but I have argued that even a non-normative thesis turns the compatibility debate into a seemingly intractable (broadly) empirical debate. Ultimately, the upshot of our discussion is that one key aspect of the "Strawsonian" program falls short of securing the non-skeptical conclusion many hoped it would secure. We can grant the thought that the standards operative in our practice cannot be mistaken – but this alone leaves it open that no one has ever met the standards that in fact operate in our practice, just as many think that no being has ever met the standards that in fact operate in *religious* practice. Ruling out this possibility seemingly requires re-engaging in all of the familiar "metaphysical" debates about moral responsibility many are understandably so keen to avoid. But those debates, unfortunately, cannot be avoided.[25]

References

Balaguer, Mark. 2023. "Strawson, Ordinary Language, and the Priority of Holding Responsible Over Being Responsible," *Harvard Review of Philosophy*. https://doi.org/10.5840/harvardreview20239753

Beglin, David. 2018. "Responsibility, Libertarians, and the "Facts as We Know Them": A Concern-Based Construal of Strawson's Reversal," *Ethics* 128 (3): 612-625.

Brink, David. 2021. *Fair Opportunity and Responsibility*. Oxford: Oxford University Press.

D'Arms, Justin, and Jacobson, Daniel. 2000. "The Moralistic Fallacy: On the "Appropriateness" of Emotions," *Philosophy and Phenomenological Research* 61: 65 – 90.

D'Arms, Justin, and Jacobson, Daniel. 2006. "Sensibility Theory and Projectivism." In David Copp, ed., *The Oxford Handbook of Ethical Theory* (Oxford: Oxford University Press), pp. 186-218.

Darwall, Stephen. 2021. "*Freedom, Resentment, and the Metaphysics of Morals*, by Pamela Hieronymi," *European Journal of Philosophy* 29: 528 - 532.

De Mesel, Benjamin and Sybren Heyndels. 2019. "The facts and practices of moral responsibility," *Pacific Philosophical Quarterly* 100: 790 - 811.

De Mesel, Benjamin. 2022(a). "Being and Holding Responsible: Reconciling the Disputants through a Meaning-Based Strawsonian Account," *Philosophical Studies* 179: 1893 – 1913.

De Mesel, Benjamin. 2022(b). "Taking the Straight Path. P.F. Strawson's Later Work on Freedom and Responsibility," *Philosophers' Imprint* 22: 1 – 17.

Enoch, David. 2005. "Why Idealize?" *Ethics* 115: 759 – 787.

Fischer, John Martin and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

[25] For helpful comments/discussion on previous drafts of this paper, I would like to thank Andrew Bailey, Neal Tognazzini, Randy Clarke, Leo Menges, Mark Balaguer, Scott Hill, and audiences at departmental seminars at Lund, Stockholm, Uppsala, Southern Methodist, UT Austin, and Innsbruck. Finally, I am grateful to Yuval Avnur for arranging this volume, and to John Martin Fischer – both for helpful comments on this paper, and for his justly influential work on moral responsibility over the years, from which I have learned so much.

Fischer, John Martin. 2014. "Peter Strawson and the Facts of Agency," in Shoemaker and Tognazzini (eds.), *Oxford Studies in Agency and Responsibility* (vol. 2). Oxford: Oxford University Press. pp. 93 – 116.

Fischer, John Martin. 2023. "The Resilience of Moral Responsibility," in Cyr, Law, and Tognazzini (eds.), *Freedom, Responsibility, and Value: Essays in Honor of John Martin Fischer.* New York: Routledge.

Fricke, Christel. 2015. "Questioning the Importance of Being Normal. An Inquiry into the Normative Constraints of Normality," *Journal of Value Inquiry* 49: 691-713.

Heyndels, Sybren and Benjamin de Mesel. 2018. "On Shoemaker's Response-Dependent Theory of Responsibility," *Dialectica* 72: 445 – 451.

Hieronymi, Pamela. 2020. *Freedom, Resentment, and the Metaphysics of Morals*. Princeton: Princeton University Press.

Holton, Richard. 1991. "Intentions, Response-Dependence, and Immunity From Error," in *Response Dependent Concepts*. Canberra: ANU Working Papers in Philosophy

Holton, Richard. 1992. "Response-Dependence and Infallibility," *Analysis* 52: 180 – 184.

Johnston, Mark. 1989. "Dispositional Theories of Value," *Proceedings of the Aristotelian Society* 63.

López de Sa, Dan. 2013. "Rigid vs Flexible Response-Dependent Properties," in Hoeltje, Schnieder, and Steinberg (eds.), *Varieties of Dependence* (Philosophia Verlag).

McGeer, Victoria. 2019. "Scaffolding Agency: A Proleptic Account of the Reactive Attitudes," *European Journal of Philosophy* 27: 301 – 327.

Menges, Leonhard. 2017. "Grounding Responsibility in Appropriate Blame," *American Philosophical Quarterly* 54: 15 – 24.

Pereboom, Derk. 2014. *Free Will, Agency, and Meaning in Life.* Oxford: Oxford University Press.

Pettit, Philip. 1991. "Realism and Response-Dependence," M*ind* 100: 587 – 626.

Russell, Paul. 2021. "Responsibility Skepticism and Strawson's Naturalism: Review Essay on Pamela Hieronymi, *Freedom, Resentment, and the Metaphysics of Morals.*" *Ethics* 131: 754 – 776.

Shoemaker, David. 2017. "Response-dependent Responsibility; Or, a funny thing happened on the way to blame," *The Philosophical Review* 126: 481 – 527.

Shoemaker, David. 2022. "Response-Dependent Theories of Responsibility," in *The Oxford Handbook of Moral Responsibility*, eds. Nelkin and Pereboom. Oxford: Oxford University

Press.

Strawson, P.F. 1962. "Freedom and Resentment," *Proceedings of the British Academy* 48, 187–211.

Strawson, Galen. 1994. "The Impossibility of Moral Responsibility," *Philosophical Studies* 75: 5 – 24.

Todd, Patrick. 2011. "A New Approach to Manipulation Arguments," *Philosophical Studies* 152: 127 – 133.

Todd, Patrick. 2013. "Defending (a modified version of) the Zygote Argument," *Philosophical Studies* 164: 189 – 203.

Todd, Patrick. 2016. "Strawson, Moral Responsibility, and the 'Order of Explanation': An Intervention," *Ethics* 127 (1): 208-240.

Todd, Patrick. 2017. "Manipulation Arguments and the Freedom to do Otherwise," *Philosophy and Phenomenological Research* 95: 395 – 407.

Todd, Patrick. 2019. "The Replication Argument for Incompatibilism," *Erkenntnis* 84: 1341 – 1359.

Todd, Patrick and Brian Rabern. 2023. "Resisting the Epistemic Argument for Compatibilism," *Philosophical Studies* 180: 1743 – 1767.

Todd, Patrick. 2024. "The Consequences of Incompatibilism," in M. Kiener, ed., *The Routledge Handbook of the Philosophy of Responsibility.* Routledge.

Van Inwagen, Peter. 1983. *An Essay on Free Will.* Oxford: Oxford University Press.

Watson, Gary. 1987. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In Schoeman, Ferdinand, ed. 1987. *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge: Cambridge University Press, 256-86.

Watson, Gary. 2014. "Peter Strawson on Responsibility and Sociality," *Oxford Studies in Agency and Responsibility volume 2*, eds. David Shoemaker and Neal Tognazzini. Oxford: Oxford University Press.

Wright, Crispin. 1989. "Wittgenstein's Rule-following Considerations and the Central Project of Theoretical Linguistics," in A. George (ed.) *Reflections on Chomsky*. Oxford: Basil Blackwell, pp. 233-64.

Wright, Crispin. 2003. *Saving the Differences*. Cambridge, MA: Harvard University Press.