

The Explanatory Component of Moral Responsibility

Gunnar Björnsson

Karl Persson

University of Gothenburg

Introduction

People who have thought long and hard about moral responsibility are still in deep disagreement. While some feel strongly that determinism undermines responsibility,¹ others think that what is relevant is how that action relates to the agent at the time of choice, not how the agent came to be such that she chose the way she did.² And many disagree whether luck of various kinds is compatible with full moral responsibility and to what extent actions must be fully determined by rational deliberation.³

As we will see, some of the most important arguments supplied in these controversies are effective insofar as they lead us to *focus* on one aspect of the cases discussed at the expense of others: to focus on the agent's motivation and deliberation as a cause of the action, or to focus on elements of luck or the existence of prior causes. Some of these arguments tend to provoke skepticism about moral responsibility as they elicit intuitions undermining our ordinary ascriptions of responsibility; other arguments have the opposite effect.⁴

The fact that changes of focus affect intuitions of responsibility raises questions: On what factors *should* we focus our attention? What focus makes for *reliable* intuitions? It is not clear that such questions can be answered by providing further cases or thought experiments, as reactions to such cases are likely to display the same sort of focus dependence.

¹ Chisholm (2003), van Inwagen (1983), Kane (1996), O'Connor (2000), Pereboom (2001)

² Frankfurt (1969), (1971), Wallace (1994), Persson (2005)

³ Nagel (1979), Strawson (1986) and Smilansky (2000) are among those who stress the responsibility-undermining effects of luck. Dennett (1984), (2003); Wolf (1990); Mele (1995), (2006) as well as Fischer & Ravizza (1998) think that those problems are surmountable.

⁴ For effects on "folk" intuitions, see Nichols and Knobe (2007), Nahmias, Coates and Kvaran (2007), e.g.

This paper approaches the problem from a new angle. It would be easier to determine what to think about moral responsibility if we were clearer about *why* we react the way we do to these arguments, and *why* our reactions vary. To this end, we will do three things.

First, we will present and motivate a general model of our judgments of moral responsibility, a model according to which such judgments are, essentially, *explanatory* judgments.

Second, we will explain how this model can account for not only factors that affect the degrees to which we assign moral responsibility in ordinary life, but also the sometimes contradictory judgments that people make in response to two of the most important skeptical arguments in the philosophical debate. Put briefly, the model can account for these phenomena because explanatory judgments are relative to explanatory interests and perspectives, and because explanatory perspectives are affected by changes in focus.

Finally, we suggest that this has important methodological consequences for the debate about moral responsibility, by undermining the assumption that apparently fundamental disagreements about moral responsibility are real disagreement of fact. Since judgments of moral responsibility are explanatory judgments, and since explanatory judgments are relative to explanatory perspectives, it would seem that apparently conflicting judgments of responsibility might all be correct if from with different explanatory perspectives. Moreover, once the question concerns what explanatory perspective we *should* take in contexts where our judgments of responsibility govern our reactive attitudes, we will suggest that whereas there are reasons to take the explanatory perspectives of everyday attributions of responsibility, there is little reason, if any, to take the perspective induced by skeptical philosophical arguments.⁵

The three components of moral responsibility

There are two reasons to take the model that we will propose here seriously. The most important reason, ultimately, and the one that we will focus on in this paper, is that the model explains both everyday judgments and judgments made in response to philosophical arguments. The other reason is that we can expect something like this model to be correct

given the role that judgments of moral responsibility play in our lives. This reason is more speculative, but since it both motivates and introduces the model, it merits a sketch.

It is well known that judgments of moral responsibility govern moral reactive attitudes. Whether someone is subject to moral requirements seem to depend on whether that person would be *responsible* for adhering to or violating those requirements. Most people think that it is appropriate to blame a person for doing something, or feel indignation towards her for doing it, or for her to feel guilt for doing it, only to the extent that she is morally responsible for it. Only then does she *deserve* being at the receiving end of these attitudes. Similarly, people think that someone is worthy of moral admiration and moral praise for an action, or can be forgiven for an action, only to the extent that she is morally responsible for it.

In fact, judgments of moral responsibility seem to have *no other* general important function in our lives than governing moral reactive attitudes and behaviors. This is not to deny that we can be said to be responsible for things that are neither praise- nor blameworthy, such as our exact choice of route as we stroll through a park, or an intricate pattern made with a stick in the sand. But talk about responsibility in general and moral responsibility in particular is most naturally situated in contexts where what someone is responsible for is something of positive or negative importance, where we care about the outcome, and where we are at least disposed to either blame or praise.⁶

Given this, it is natural to think that our concept of moral responsibility, or our capacity to identify cases of moral responsibility, has been shaped by the need to pick out suitable objects for reactive attitudes and behaviors. By “suitable objects”, we do not mean the objects towards which such reactions are, ultimately, justified, but rather objects the attitudinal reactions towards which tend to be reinforced.

⁵ Based on a different set of arguments, similar claims have been made about judgments about *free will* or *free action* by John Hawthorne (2001) and Steven Rieber (2006).

⁶ The locus classicus on the importance of the connection between reactive attitudes and moral responsibility is of course Sir Peter Strawson’s seminal (1974) paper “Freedom and Resentment”.

As Angela Smith (2007) stresses, moral responsibility is only one among several factors that determine whether an agent should be blamed or taken to be culpable for something. The degree of fault involved matters, as well as the agent’s own response to what has been done, and blame and the expression of blame might only be appropriate if one stands in the right relation to the responsible agent.

It is also natural to think that such reinforcement is largely determined by whether the reactions help to further the obvious social and psychological functions of moral reactive attitudes. Such attitudes are used, consciously and unconsciously, to control and shape motivational structures – priorities, values, preferences, desires, behavioral and emotional habits, etc. – *in order to promote and prevent certain kinds of behaviors or events* (typically through causing reflection on and perhaps re-evaluation of values involved). We might of course be indignant with someone's lack of concern or feel guilty for having certain feelings even if neither has resulted in inappropriate behavior, but even when the attitudes are directed at motivational structures, it seems clear that what upsets or pleases us about them is that they are dispositions to act in certain ways. Moreover, moral reactive attitudes are most often directed straight at behaviors understood as governed by such motivational structures: at ourselves or others for failing to do something for lack of compassion, say, or for doing something out of greed.

In order for our reactive attitudes to properly control and shape motivational structures and promote and prevent various behaviors or events, we need to direct them towards the sort of motivational structures that (a) explain these events in systematic ways and (b) respond to reactive attitudes in the appropriate way. The suggestion, then, is that mechanisms that tend to direct our attitudes towards such motivational structures are reinforced. But we have also argued that the main function of our concept of moral responsibility is to govern these attitudes. Our hypothesis, therefore, is that people take P to be morally responsible for E to the extent that they take E to be explained in normal ways by some motivational structure, S, of P that is of a kind that can be modified by reactive attitudes and reflection on the values involved.⁷ Call this the *Explanation Hypothesis*.

More precisely, on this way of thinking, P is responsible for E to the extent that GET, RR and ER are satisfied:

⁷ Although the account offered here allows for responsibility for outcomes of joint efforts as well as collective responsibility (assuming, as seems plausible, that collective agents can have motivational structures that respond to reactive attitudes), our concern here is with the responsibility of individuals when these are not understood as part of a larger group. Moreover, although the account allows for responsibility for one's own motivation and, to some extent, beliefs, our concern here will be with responsibility for actions and external events.

General Explanatory Tendency (GET): There is a reasonably common condition C such that motivational structures of type M explain outcomes (actions, events) of type O given C while motivational structures of type M' explain outcomes of type O' given C, where M and M' as well as O and O' are mutually exclusive.

Reactive Response-ability (RR): Generally speaking, whether people exemplify M or M' depends on what sort of reactive attitudes (if any) they are subject to for realizing O or O' or on whether they reflect on the kinds of value realized by O and O'.

Explanatory Responsibility (ER): P has a motivational structure, S, of type M; E is of type O; C holds; and S explains E given C in the normal way that M-motivation explains O-outcomes given C.

Intuitively, GET demands that M should make a systematic difference for a certain kind of outcome, RR demands that M is subject to modification in light of its tendency to make this difference, and ER demands that the case in question instantiates the right sort of explanatory pattern.

Obviously, the Explanation Hypothesis is not meant to capture everything people say about moral responsibility. In natural language, the sense of ordinary expressions is modified by communicative needs of particular contexts, and there are a number of different senses of “responsibility”. For example, there is an institutional or social sense of responsibility according to which we take on and distribute areas of responsibility (“Wilma is responsible for bringing wine, Fred for bringing food”; “The Chancellor of the Exchequer is responsible for the budget”), as well as a wide causal or explanatory sense of responsibility according to which the question “what is responsible for E?” is interchangeable for “why did E happen?”⁸. Moreover, there is a practice of directing reactive attitudes towards people – “holding them responsible” or “taking them to task” – not only for things they bring about or fail to prevent, but also for things they endorse, or things brought about by what they endorse; such cases typically violate ER.⁹

⁸ For discussion of relations between some notions of responsibility, see Watson (1996).

⁹ This practice might explain why Woolfolk et al (2006) found that when an agent desired the death of a friend, he was taken to be somewhat responsible for killing his friend even in cases of extreme coercion where the agent had been forced to take a “compliance drug” and ordered to commit the murder.

Nevertheless, we suggest that most of our strongest intuitions about moral responsibility, and the reactions that are very much driving the philosophical debate, are tied to GET, RR and ER.

The focus of this paper is on ER, but a few words are needed to avoid misunderstanding of GET and RR. Both these conditions are meant to capture the idea that our concept of moral responsibility tracks *kinds* of motivation towards which it is worth directing reactive attitudes because it explains salient types of outcomes *in general*, not just in special cases.

In particular, GET rules out that we are responsible for everything that happens to be explained by our motivational states for some unusual reason. For example, if Mr. Black starts stalking Mr. Grey because Mr. Grey likes to wear plaid vests and because Mr. Black is obsessed with plaid vests, Mr. Grey is not thereby responsible for being stalked by Mr. Black. Since it is not *in general* the case that liking something explains being stalked, this case violates GET.

The explanatory relations that most clearly do satisfy GET involve explanations of outcomes with reference to preferences that certain those outcomes take place, or lack of preference that they do not take place. Accordingly, and *ceteris paribus*, we take people to be responsible for outcomes that are explained in normal ways by their desires for such outcomes, and to be responsible for outcomes that are explained in normal ways by their lack of sufficient concern for such outcomes.¹⁰

To understand RR, it is important to keep in mind that it concerns motivational structures of certain *types*, not the instantiation of motivational structures in a particular individual at a certain time. When a reckless driver dies as he crashes into another car, we might hold him responsible even though his death obviously prevents any further changes to his motivational structure; what is required is that we understand the *type* of motivational structure that explained his reckless driving and the accident to be responsive in the right way. By contrast, RR is undermined in the case of overwhelming motivational states, when the agent has a

¹⁰ Typically, such explanations demand that the agent *knows* about a possible outcome and about available means to prevent or promote it. But motivation or lack thereof also normally explains outcomes by explaining why we notice or fail to notice certain possible outcomes, as when a negligent father misses signs that his children needs him because he cares more about his work. Even though there is a sense in which such outcomes are beyond our control as agents, we readily attribute responsibility. See Sher (2006).

general incapacity to respond with self-directed reactive moral attitudes such as guilt and shame, or when she lacks self-control in general, or the capability to predict the consequences of one's actions. RR explains why we take moral responsibility to be diminished by phobias, compulsive behaviors, severe anxiety, psychopathy, autism and serious personality disorders. For that reason, something like RR is an integral part of most compatibilist accounts of moral responsibility.

Enough has been said to provide a rough idea about the content of GET and RR. In what follows, we will focus almost exclusively on ER, which explicitly concerns the particular event for which we hold someone responsible.

ER is most clearly violated in cases of overwhelming external obstacles. If external obstacles make it impossible for me to do something, the fact that I am not doing it cannot normally be explained with reference to my motivational structure, (unless I am responsible for the external obstacles). Consequently, I am not seen as morally responsible for not doing it.

In one form or other, requirements like these are part and parcel of most compatibilist accounts. ER has much more interesting consequences, however, having to do with the fact that the notion of *what explains E* is *selective* or *interest-relative* in a certain way. What we will try to show is how this has significant consequences for our everyday notion of moral responsibility.

Selective explanatory interests and everyday excuses

Ordinarily, when we are looking for the causal explanation of some event or condition, E, we are not trying to assemble any or all conditions or events that can be said to make a causal contribution to the occurrence of that event; we are not asking for a complete and maximally detailed description of its causal origins, or a complete explanation of why it came about. We are trying to identify a condition that is especially *relevant* given our explanatory interests.

Typically, such a condition, X, only provides a causal explanation of E given a number of further conditions, C, which we might call the *supporting conditions* of X's explaining E. Nevertheless, as we think that X explains E, our focus is on X and E, while C is part of the cognitive background of our thought; cognitively, X and E are treated as variables, while C is treated as a constant.

One of the factors that determine whether X is an *interesting* aspect of what explains E is whether X is more remarkable, surprising or out of the ordinary than the background

conditions. When the smoke detector sounds its alarm, a complete causal explanation of the event will include various facts about the wiring of the detector, the fact that it has a good battery, and the presence of smoke. However, given that we expect the detector to be in good working order, what we would think of as *explaining the alarm* is the presence of smoke. If we had expected the presence of smoke but not that of the battery, we would have thought of the latter condition as what explained the alarm.

The interest-relativity of everyday explanatory judgments is well known, but has surprising explanatory power when ER is understood as selective and interest relative in this way.¹¹ Consider the force and limits of six kinds of everyday excuses, or considerations that lower moral responsibility:

1. *He was forced to do it.* The Explanation Hypothesis explains why various degrees of external force, threats and constraints reduce moral responsibility to corresponding degrees: these factors reduce the *explanatory relevance* of the motivational structure of the agent. This is clearest when someone else moves an agent's limbs against his will, or physically stops a person from performing an action that he wants to perform. The person's motivational structure fails to explain both the person's movement in the first case, and the fact that he did not perform the action in the second case; as ER predicts, we take him to be responsible for neither.

The same sort of reduction of responsibility occurs, though less obviously, when someone imposes great costs on certain types of action from the outside, threatening to destroy or hurt what the agent values. As the threats grow more extreme, the agent's motivational structure becomes less interesting in explaining the outcomes of his actions, because almost any normal motivational structure would yield the same action.¹² Compare two cases where a bank clerk hands over the money to robbers. In the first case, the robbers had threatened to dump rotten vegetables the windshield of the clerk's car; in the second they had threatened to kill the clerk and some customers. Why did the bank lose its money? In the first case, it would make sense

¹¹ The reader will recognize similarities between the way we apply this idea to judgments of moral responsibility, and the way it was spelled out by Hart and Honoré (1985: 33-44) and applied to legal responsibility.

¹² If this isn't obvious enough, it is confirmed by Woolfolk et. al. (2006), who report experiments where increasing coercion shifted assigned responsibility from agent to coercer.

to mention the motivational structure and decision of the clerk, but hardly in the second, since almost anyone would have acted in that way. Consequently, there is significant reduction of the clerk's responsibility for the bank's loss only in the second case.

Notice that what we are explaining here is a *tendency*. It is *possible* to think of the second case in ways that take the clerk to be responsible for the loss; in fact, the clerk might take herself to be so responsible, and also, by the same token, responsible for the lives saved. What is needed to achieve this effect is that we can envision alternative motivational structures that satisfy RR and GET, and that the threat becomes part of the background against which the loss is to be explained. One way to achieve the relevant backgrounding is to encourage taking up the clerk's perspective of choice, where the circumstances – including the threat – are given.¹³ However, since the bank's loss is considerably more surprising given the clerk's motivational structure than given the robber's threat, the question of why the bank lost the money is, generally speaking, more likely to be framed against a background which doesn't contain the robber's threat, but does contain the clerk's motivational structure. From this perspective, the clerk will not seem responsible for the loss.

2. *It wasn't under her control.* A driver suffers a brain hemorrhage while on the highway. As a result, her reflexes are impaired, and her car crashes into the breaking car in front. Knowing this, we don't hold her responsible for the crash. ER explains this nicely: the hemorrhage, not the driver's motivational structure, explains the accident. Moreover, ER also explains why we often *do* hold agents responsible for outcomes that are not under their control at the time. If the driver started driving knowing that she would have reduced control over the vehicle – perhaps because she had been drinking, or taking medication, or been deprived of sleep – we are likely to hold her responsible for an accident even if her failure to avoid it was not due to any lack of motivation *when it happened*. In such a case, a proximate explanation of the crash would perhaps focus on her remarkably slow reflexes. But a more distant explanation of this feature of the situation would also stand out: her decision to drive with reduced control. And this decision, and the motivation behind it, are remarkable given the risks involved, and given normative and legal expectations to drive responsibly.

3. *He just did his job vs. he broke the rules.* Compare two cases in which the receptionist at the clinic unlocks the door for someone who is carrying a child in need of emergency care, and in which doctors manage to save the child's life. In both cases, the child would have died had the receptionist not unlocked the door, but whereas the receptionist in the first case is merely doing his job, the receptionist in the second deliberately violates strict orders not to let unauthorized personnel into the building. When we want to explain *why* the child was saved, we are likely to mention the receptionist's action in the second case but not in the first. As ER predicts, we are also more prone to take the receptionist to be responsible for the child's survival in the second case.

We see the same effects when the outcome is unintended and negative. The receptionist at an apartment building lets burglars into the building, and several tenants are burglarized. In one case, the receptionist suspects that the people that are let in might have illegitimate business, but unlocks the door as standard practice is to refuse entry only to *known* troublemakers. In the second, the suspicion is the same, but the receptionist unlocks the door *even though the rules explicitly say that only tenants should be let in*. Again, we assign a much higher degree of responsibility in the second case, where the willingness to unlock the door would be seen as relevantly explaining why the tenants got burglarized.¹⁴

4. *He didn't know that it would happen.* In general, we think that people are less responsible for an event if they bring it about or let it happen unwittingly than if they do it knowingly. This is straightforwardly explained by ER. Suppose that a hoodlum steals a lady's purse

¹³ Given this perspective, the agent rather than the circumstances will seem responsible for whatever choice is made. This, then, is a perspective in which radical existentialist claims about responsibility will seem reasonable; the Explanation Hypothesis explains why such claims doesn't always seem completely unrealistic.

¹⁴ We take it that normative expectations affect explanatory interests by affecting allocations of explanatory conditions to foreground and background. This is clearly illustrated in a study by Alicke (1992), which suggested that perceived culpability affects what we take to be the "primary" cause of an event, and in a recent study by Joshua Knobe and Ben Fraser (2008). Knobe and Fraser presented a scenario in which faculty members were not allowed to take pens while the administrative assistants were. Usually, the faculty members disregarded the prohibition and took pens anyway. One day, a professor took a pen and an assistant took a pen. Later during the day one of the assistant needed a pen to write down an important message but the pens were gone. When subjects were asked who caused the problem, a vast majority claimed that it was the professor even though both the professor's and the assistant's action were the same.

behind the back of a police officer, Pete, and that Pete did not hear what was happening. Because of this lack of knowledge, it seems that we cannot explain the hoodlum's successful robbery with reference to Pete's motivational structure. Whatever his motivation, he would have done nothing about the robbery because he knew nothing about it. And, as ER predicts, we are now unwilling to assign moral responsibility to Pete.

ER also predicts that moral responsibility sometimes survives ignorance. Suppose that Pete failed to notice the robbery because he was busy listening to the dog racing results on the radio while on patrol duty. Now his ignorance seems to be the result of a remarkable disregard for his job: *Why was the robbery successful? Because Pete found the dog racing results more important than the street life!* As ER predicts, Pete is now found morally responsible for the lady's loss – although less so than the hoodlum.

5. *She didn't do it, she was just a bystander.* ER neatly explains why we tend to hold people who have actively and intentionally produced an outcome to be more responsible for it than people who have been mere bystanders but could have intervened. Suppose that when the hoodlum yanked the purse from lady's hand, Linda saw what happened and could have stopped the hoodlum. Nevertheless, if asked why the lady no longer has her purse, we will naturally focus first on the intentional action of the hoodlum, and only later on Linda's inaction. Consequently, we will primarily assign responsibility for the lady's loss to the hoodlum.

Being a witness rather than actively pursuing E does not always remove moral responsibility completely, and ER explains that too. According to ER, Linda's inaction could make her responsible if it were *remarkable* that she lacked the motivation to intervene. Suppose that, like Pete, Linda is a police officer; that intervening would have been a simple matter; and that she decided not to intervene because it would ruin her coffee break. Given this information, it would seem reasonable to say that the lady lost her purse *because Linda cared more about her coffee break than protecting the public.* And just as ER predicts, Linda now seems to be morally responsible for the lady's loss. Suppose instead that Linda is an ordinary civilian, both smaller and weaker than the hoodlum, and as likely to get hurt as to retrieve the purse. Now her lack of willingness to act is unremarkable and unsuitable as an answer to the question of why the lady lost her purse. Consequently, we now judge her moral responsibility for the loss of the purse as nonexistent, or at least radically lower than in the officer case. Again, this accords with ER.

In ordinary thought, there seems to be a tension between holding the failing officer responsible for the lady's loss and holding the hoodlum responsible. At the same time, most people who think about cases like this seem to agree that both can be responsible, although for different reasons and in different ways. ER explains both the tension and the compatibility. The two assignments are in tension, because while focusing on the fact that one action explains the outcome, we will tend to treat the other action as part of the explanatory background condition, and thus, for the moment, as unremarkable and non-explanatory. At the same time, they *are* compatible because we can either shift between the two explanatory frames, or widen our view and take the conjunction of the two actions to be what explains the loss. However, this doesn't completely remove the tension, because taking them to be compatible is cognitively much more complex, forcing us to take a more abstract view of the matter.

6. *It wasn't her initiative.* Finally, ER neatly explains why someone who is actively pursuing an end and engages others in the effort tend to be seen as more responsible for achieving it than those that are being engaged. Sarah, an ordinary civilian, manages to stop the hoodlum from our previous example, but has difficulty controlling him and calls for bystanders to help. Catherine, another civilian, is the first to answer the call, and together they get the hoodlum pinned to the ground. It is natural to say that the lady gets her purse back because of Sarah's and Catherine's willingness to help. According to ER, then, it is natural to take both to be morally responsible for getting the purse back. But Sarah seems *more* responsible, and ER explains that too. Not only is it more remarkable that someone is willing to take the initiative to stop a crime than to answer a call to join someone who has shown the way, but we also take Sarah's action to explain Catherine's, and thus to be explanatorily more basic. (Notice, though, that this tendency can be counteracted. Suppose, for example, that Sarah is a police officer who is just doing her job, while Catherine is a civilian: now Catherine and Sarah might seem more equally responsible for getting the purse back.)

The Explanation Hypothesis would seem to gain considerable credibility from its capacity to predict what our judgments of moral responsibility will be in these cases. It is a striking fact that our judgments – both positive and negative – seem very well matched by corresponding explanatory judgments. Someone might worry, though, that the reason for this coincidence is that explanatory judgments are influenced by responsibility judgments and judgments of blame- and praiseworthiness, rather than the other way around. And this worry might be

bolstered by the fact, mentioned above, that we are especially prone to pick out as causes or explanatory events those that are norm-transgressions. However, the evidence for the Explanation Hypothesis doesn't only consist in the coincidence of positive and negative explanatory judgments and corresponding responsibility judgments. In general, ordinary explanatory judgments are sensitive to whether something is, given normal expectations, a *remarkable* part of a complete explanation, and we have provided reasons, in each case, for expecting various motivational structures to be remarkable, or not. If our discussions of the various cases have been on track, this means that there are independent reasons to expect the explanatory judgments in question. And this, in turn, means that there are independent reasons to accept the Explanation Hypothesis.

Explanatory perspectives on heteronomy: regress arguments

We have seen how the Explanation Hypothesis explains a number of common sense intuitions about moral responsibility: whether someone is taken to be responsible for an event seems to depend on whether some RR-satisfying motivation is an especially *relevant* part of a complete explanation of that event, relative to our explanatory interests. In what follows we will see how the pragmatics of explanation is equally capable of explaining central philosophically puzzling aspects of our thinking about responsibility. In this section, we focus on the most popular arguments for skepticism about moral responsibility, arguments from manipulation and heteronomy; after that we will do the same for a problem with luck that has been raised for forms of libertarianism about free will and responsibility. In both these cases, we will show how the arguments that seem to undermine responsibility do so by manipulating what is naturally taken as interesting explanatory features and what is taken as background. In the final section, we will suggest that there is no reason to go into the rather special explanatory frames induced by these arguments, but good reasons not to. In effect, then, our model of how our judgments of responsibility are formed provides a defense of ordinary ascriptions of responsibility.

Skeptical arguments against moral responsibility typically appeal to what might loosely be called "heteronomy": by the fact that our actions are ultimately determined, to the extent that they are determined, by factors for which we are clearly not morally responsible. Here is one such argument, from Galen Strawson (1994: 7):

- (1) It is undeniable that one is the way one is, initially, as a result of heredity and early experience, and it is undeniable that these are things for which one cannot be

held to be in any way responsible (morally or otherwise). (2) One cannot at any later stage of life hope to accede to true moral responsibility for the way one is by trying to change the way one already is as a result of heredity and previous experience. For (3) both the particular way in which one is moved to try to change oneself, and the degree of one's success in one's attempt at change, will be determined by how one already is as a result of heredity and previous experience. And (4) any further changes that one can bring about only after one has brought about certain initial changes will in turn be determined, via the initial changes, by heredity and previous experience. (5) This may not be the whole story, for it may be that some changes in the way one is are traceable not to heredity and experience but to the influence of indeterministic or random factors. But it is absurd to suppose that indeterministic or random factors, for which one is *ex hypothesi* in no way responsible, can in themselves contribute in any way to one's being truly morally responsible for how one is.¹⁵

If we add that one cannot be responsible for actions unless one is responsible for the aspects of oneself from which these actions result, the upshot is that one cannot be morally responsible for one's actions.

We agree with Strawson that this argument and others like it tend to have great intuitive force; what we will do here is to show how the Explanation Hypothesis can account for this force. The general suggestion will be that these arguments tend to change the explanatory frame – the expectations and explanatory interests – within which we consider an agent's actions. This, in turn, changes whether we take reference to the agent's motivation to provide good explanations of these actions, and so, given ER, changes our judgments of moral responsibility.

To understand how this works, we need to have a clearer understanding of why we say or think of one particular event E as *the explanation* of another event E', when we know that E is just one event in a long causal chain leading up to E'. For illustration, suppose that we knew the following:

¹⁵ See also van Inwagen (1983, 2000), Strawson (1986), Kane (1996) and Pereboom (2001).

Sam arrived half an hour late for a meeting. One driver had been using her mobile phone, while another was having an argument with his wife; both were slow to react to changes in traffic and bumped into each other. One thing led to another, and a number of cars crashed hard into the cars in front, blocking three out of four lanes on the highway for over an hour. Sam spent almost an hour behind slow-moving cars that were stuck behind other slow-moving cars, ... , making their way past the site of the accident. Five minutes before the meeting, Sam still had 15 miles to go. Naturally, she couldn't get here in time.

Suppose further that someone asks us why Sam arrived late, and that we have time for a one-liner as we are leaving in a hurry. Compare the following answers, all of which picks out a condition that is part of the complete causal history of Sam's late arrival:

- (a) Five minutes before the meeting, Sam was 15 miles away.
- (b) Sam got stuck behind slow-moving cars for almost an hour.
- (c) All lanes but one were blocked for over an hour.
- (d) There had been a road accident.
- (e) Someone had an argument with his wife on the highway.
- (f) Someone used the mobile phone while driving.

We are guessing that for someone with statistically normal expectations among westerners, answer (a) would immediately raise the question of *why* Sam was 15 miles away five minutes before the meeting. Although providing a condition given which the late arrival could very much be expected, it isn't the *kind* of answer one would normally be interested in; one would want to know something about her as an agent – her decisions, motivation, beliefs – or about what *happened* to her, such that the late arrival could be expected. Answer (b) would raise the question of why the cars were driving slowly for such a long time and (c) would raise the question of why the lanes had been blocked; although both events make a late arrival likely, they call for further explanation because neither is the sort of thing that happens without some one straightforward explanation (road work, parade, accident), an explanation that would itself provide a straightforward explanation of Sam's late arrival. Answers (e) and (f) are defective in a different way: it is not part of the explanatory background that arguments and mobile phone use lead to late arrivals, so something more needs to be said. In contrast to all these, (d) would answer the question without either raising further explanatory questions or forcing the hearer to do a lot of guessing. We are typically taking for granted that an accident

is the sort of thing that just happens, for a variety of reasons, and the sort of thing that delays people. For these reasons, we most naturally explain Sam's late arrival with reference to the accident.

In this case, an explanatory "regress" is blocked by the fact that invoking prior causes of the accident would unduly complicate the explanation; although we get a more satisfying explanation by moving from (a), (b) or (c) to (d), no such gain is had by moving further back to (e) or (f). The same is true about everyday explanations of events in terms of motivational structures. First, such explanations take place against various background assumptions of (*ceteris paribus*) explanatory connections between RR and GET-satisfying motivational structures and the explanandum. For that reason, explanations in terms of such structures will often be unlike (e) and (f) above; we would take the fact that John cares very little about his dog to straightforwardly explain – explain without raising further questions – why the dog hasn't been well fed, and we would take the fact that Jane likes Beethoven and knew that his Fifth Symphony would be played at the concert to straightforwardly explain Jane attendance. Second, we often have no expectations of straightforward explanations of specific RR and GET-satisfying motivational structures. For that reason, explanations in terms of such structures are often unlike (b) and (c); we probably wouldn't expect straightforward explanations of why someone doesn't care about his dog, or likes Beethoven (or at least none that doesn't invoke some other RR and GET-satisfying motivational structure that itself lacks such an explanation).¹⁶ Third, we are very often interested in just the sort of answers offered by these explanations; when we are, they are unlike (a). For these reasons, reference to motivational structures will often be thought of as explaining the action or event in question, just as (d) is naturally seen as the reason for Sam's late arrival.

From a concrete and commonsensical perspective, then, desires and other motivational structures that satisfy GET and RR often do explain particular actions. Given ER, that accounts for the fact that we often do attribute responsibility. When we are confronted with regress arguments against moral responsibility, however, we are led to *abstract away* from the particulars and think in terms of "prior causes", or "heredity and early experience" and

¹⁶ Hart and Honoré (1985: 73-76) convincingly argue that when we trace consequences of actions, we stop when these consequences result through too unlikely a coincidence of independent factors. Similarly, we stop tracing prior explanations of an event when that explanation relies on too unlikely a coincidence.

“indeterministic factors”. This removes the explanatory significance of motivational structures in two steps.

To begin with, it eradicates differences in perceived complexity between, on the hand, explanations of the action or outcome in terms of the agent’s motivational structure and, on the other, explanations in terms of what in turn explains this structure. Abstract talk about *prior causes* or *heredity and prior experience* provides summaries of what would otherwise be understood as enormously complex sets of prior conditions and explanatory relations. This means that there is no longer any additional cognitive cost involved in thinking of the explanandum as being explained by heredity and prior experience rather than as being explained by the agent’s motivational structure, in the way that there was an additional cost involved in thinking of Sam’s late arrival as being explained by (e) or (f). Furthermore, since these prior causes are mentioned explicitly, we now expect motivational structures to have straightforward explainers (*heredity, early experience* and *indeterministic factors*), much like (b) and (c).

The upshot is an explanatory regress of the kind that pushed us from (a), (b) and (c) to (d) in explaining Sam’s late arrival. We are now pushed to think of our actions as being explained by heredity and early experience!¹⁷ Given ER, that means that agents will not seem responsible for their actions.¹⁸

The reason that regress arguments seem compelling, we have argued, is that they affect the explanatory judgments that constitute our judgments of moral responsibility by affecting

¹⁷ Of course, the regress need not stop there if even earlier abstractly characterized causes have been made explicit.

¹⁸ It should be noted that it also allows for individual variation in that appeal, and variation from case to case, depending on how inclined people are to take the abstract perspective. It seems plausible that people will be less inclined to take the abstract perspective when confronted by cases that involve more striking moral transgressions, since such transgressions are prone to capture our cognitive and affective focus and thus to keep in place the concrete perspective under which the details that make them striking transgressions are well in sight. The explanations suggested by the Explanation Hypothesis thus seem to be supported by recent experiments by Nichols and Knobe (2007) that indicate both that abstractly and concretely characterized cases of action yield different intuitions of responsibility, and that increased moral or emotional importance of the actions described decreases the tendency to draw skeptical conclusions from the existence of sufficient prior causes. We discuss the relation of the Explanation Hypothesis to these experiments elsewhere.

explanatory frames. But this account of the *effectiveness* of regress arguments might seem to remove their *evidential value*. The reason is that the *correctness* of ordinary explanatory judgments seems to be relative to the explanatory background against which we are asking for an explanation.¹⁹ With the right explanatory background – one in which driving while using a mobile phone is understood to be the sort of thing that leads to road accidents – it might make good sense to think of the mobile phone use as explaining Sam’s late arrival, and in a context in which we have considered road accidents without serious traffic repercussions, citing the accident might not explain – help us understand – why Sam was late. If the correctness of ordinary explanatory judgments is relative to explanatory frames in this way, and if judgments of moral responsibility are explanatory judgments, then the correctness of these judgments too might be relative to explanatory frames. If so, the denials of responsibility elicited by regress arguments would not contradict positive arguments of responsibility made from a concrete everyday perspective. At most, they would show that there is *a sense* in which we are not responsible for our actions.²⁰

This is an interesting enough result in itself, and it raises the question of whether we *ought* to assess responsibility and govern our reactive attitudes from one perspective rather than another. As we will see, the same question is raised by another family of skeptical arguments that we will consider, and we will begin to address it before closing, arguing

¹⁹ One way to capture this idea is to say that the notion of a cause is contrastive: to say that A caused B is to say that A-rather-than-A' caused B, or perhaps that it caused B-rather-than-B' (see e.g. Northcott 2008). What we have called explanatory frames would be the sort of things that determine, in context, what the relevant contrasts are.

²⁰ Replies to skeptical arguments against free will according to which these arguments changes the meaning of our judgments of free will have been proposed by John Hawthorne (2001), Daniel Dennett (2003:93 e.g.) and Steven Rieber (2006). Hawthorne suggests that “X does Y freely” implies that X’s action is free from causal explainers beyond S’s control *apart from those causal explainers that we are properly ignoring*, where what we are properly ignoring depends on context. Dennett argues that judgments made from the God’s eye perspective are different than everyday arguments, and irrelevant to the sort of freedom that we ought to be concerned with. Rieber, finally, suggests that “X does Y freely” means “X caused Y and was the original cause in doing so”, where what is taken as “the original cause” depends on what causes are salient in a context. Rieber’s proposal might be closest to ours (although he explicitly refrains from drawing conclusions about moral responsibility), but ER gets more wide-ranging support by its capacity to predict a number of aspects of our thinking about moral responsibility that are unrelated to skeptical arguments and the existence of prior causes.

briefly that our reactive attitudes should be governed primarily by judgments made from the everyday perspective.

Contrastive explanations and problems of luck

Many philosophers have pointed out that luck seems to undermine moral responsibility in various ways. In his famous (1979) paper “Moral Luck”, Thomas Nagel argues that when we get a “more complete and precise account of the facts”, we understand that factors outside of our control have a great influence over our actions and their consequences, and that makes us less inclined to hold people morally responsible for what they do (Nagel 1979: 26-28). Strawson’s argument from heteronomy provides an example where our motivational structure and our actions are determined or caused by factors that precede our capacities for self-determination, thus making it a matter of luck that we are what we are and choose what we choose. In this section, our concern is the possibility that our decisions and actions are not determined by what we take to be our reasons for action; relative to our motivational structures and our rational practical thinking, what we actually do would then be, at least in part, a matter of luck. We will look at a few considerations adduced in the debate, and show how ER can explain the force of these considerations.

Imagine an agent, Alice, whose actions are not strictly determined by her stable rational judgments about what she has reason to do. This might be either because other, non-rational, factors influence her decisions, or because the connection between rational thought and decision is indeterministic. As a result, there is only a weak probabilistic connection between what she takes as reasons for action and her actual choices and actions. For example, even though she typically thinks that she has most reason to keep a promise, she will often decide to break it. It seems that Alice’s lack of rational control undermines her responsibility for at least those of her actions that do not accord with her stable judgments about what she has reason to do. Moral responsibility seems to diminish where rational thought is less of a determinant of our actions. This, of course, is one implication of the Explanation Hypothesis; Alice’s actions are not explained by the kind of motivational structures that are systematically modifiable by reactive attitudes and evaluative reflection.

Presumably, we all share Alice’s predicament to *some* degree; our actions are not always completely determined by our rational thought and control mechanisms. This might not seem to be a deep problem, as we might nevertheless have a high enough degree of rational control most of the time, and since some lack of rational control seems compatible with moral

responsibility. For example, suppose that John has considered, from time to time, the possibility of killing Bill, the owner of the big corporation that has brought his small business to bankruptcy, as a way of getting revenge. Sometimes that has seemed like a good idea; at other times like a terrible idea. Suddenly, one day, as John is hunting deer in the woods, he sees Bill standing alone in a small glade. He thinks to himself, “I really shouldn’t, but I’m gonna”, raises his rifle and kills Bill with one shot. Suppose further that John’s decision and action in this case were not completely determined by his capacity for rational control and motivational structure just before the decision. Given his wants, desires, habits, and rational evaluation of behavior, his decision *could* have gone both ways. As with Alice, this could either be because the causal connection between motivational structure and decision was sensitive to factors that seem irrelevant to John’s control over his decisions – random “noise” in his neural activity, say – or because of brute indeterminism in his decision making system. In either case, it still seems possible for John to be responsible for shooting and killing Bill, at least if John has a history of responsible behavior and decisions.

It is clear that the Explanation Hypothesis, and in particular RR and ER, makes it possible or likely that we will take John to be responsible for the killing. Unlike Alice’s motivational structure, John’s is of a kind that responds in systematic ways to reactive attitudes and reflection over outcomes. Moreover, even though John’s motivational structure was causally insufficient for the outcome, it would still seem to play a crucial explanatory role – had John been more resolutely for abiding by the law and less concerned about getting revenge, he would not have fired the shot.

Attributions of responsibility in cases like John’s are grist for the mill for libertarians who take indeterministic choices like John’s to be the source of full moral responsibility: they show that indeterminism need not seriously undermine responsibility.²¹ However, Al Mele (2005; 2006, ch. 3) has recently put the problem of luck in a new and intuitively forceful way.

²¹ Chisholm (2003), van Inwagen (1983) and Ekstrom (2000) take responsibility for an action to demand that, at the moment of choice, the choice is undetermined by any prior state of the world or agent, and that there was some alternative action such that had the agent performed it, the agent would have been responsible for that action too. Other libertarians argue that an agent can be responsible even though, at the moment of choice, there are no alternative possibilities (cf. Kane 1996, Pereboom 2001, Mele 2006: ch. 5); what is important, instead, is

Consider two worlds: the actual world, *W*, in which John decides to kill Bill, and a possible world, *W'*, which is exactly similar to the actual world up to the moment of John's decision but where John doesn't kill Bill. Applying Mele's challenge to this case, it would consist in the following argument:

1. There is no difference between John in *W* and John in *W'* that explains why John decides to kill Bill in *W*, but not in *W'*.
2. Hence, the difference in action is just a matter of luck.
3. Because of this, it is a matter of (bad) luck that John kills Bill.
4. And because of this, John cannot be responsible for his action.

Mele is not alone in feeling that these considerations *seem* to considerably undermine John's responsibility for his deed. We share that untutored intuition, and so does Randolph Clarke (2005: 416-19). And while both Mele and Clarke think that the challenge can be met, they think that it continues to carry *some* weight. To the extent that it does, libertarians of the sort that take responsibility to rest on cases where both (a) and (b) hold remain in an uneasy position.

Libertarians are not alone in that uneasy position, however. Mele's challenge seems to generalize from indeterministic cases to deterministic cases of diminished rational control:

- 1'. Among factors that are relevant to John's capacity to rationally control which decision he makes, or factors that constitute those of his motivational structures that can be modified by reactive attitudes or reflection over values, there is no difference between John in *W* and John in *W'* that explains why John decides to kill Bill in *W*, but not in *W'*.
- 2'. Hence, from the point of view of John's capacity to responsibly determine his actions, the difference in action between *W* and *W'* is just a matter of luck.
- 3'. Because of this, it is a matter of (bad) luck, relative to John's capacity for rational action, that John kills Bill.
- 4'. And because of this, John cannot be responsible for his action.

that the agent's actions have been undetermined at earlier times such that the agent can be said to have created the character or motivation that now determines the actions.

Whether determinism or indeterminism is true, this sort of luck seems to undermine responsibility. And it might be quite common. For the first premise of either of the two arguments above would be true even if John had done what he thought he had most reason to do and even if the probability were quite high that he would act according to his rational evaluation given his motivational structure and capacity for control over his decisions and actions. Call the threat posed by these arguments the problem of “contrastive” luck.

The Explanation Hypothesis accounts for why the arguments seem to threaten responsibility. The reason is that the contrastive explanatory claim of the first premise necessarily relegates factors relevant to John’s capacity to responsibly determine his action to the explanatory background conditions, thus changing the explanatory frame that was involved in coming to the prior conclusion that John is responsible. In explaining why John killed Bill in *W* but not in *W'*, we need to identify some feature that *differs* between *W* and *W'* and that makes John’s killing Bill more likely in the former. And, per hypothesis, there is no such feature among those that are relevant to John’s capacity to responsibly determine his actions. Asking for a contrastive explanation of this sort thus forces potentially explanatory factors – including John’s RR-satisfying motivational structure – into the *background conditions*. When this backgrounding is psychologically active as we turn to the question of John’s responsibility for killing Bill, we will tend to take indeterminism or features outside John’s motivational set or John’s control to be comparatively more relevant, and what is inside as less relevant. Given ER, this means that we will ascribe less or even no responsibility to John; his motivational structure does not strike us as an interesting explanatory feature.²²

²² Clarke (2005: 415) subtly misrepresents the contrastive explanation involved in Mele’s argument. The explanatory question is not why John decided to kill Bill rather than not kill Bill – why the actual world is such that John decided to kill Bill rather than not – but why John decided to kill Bill in *W* but not in *W'*. In cases where John was quite likely (but not fully causally determined) to kill Bill, the former has an answer – John wanted revenge, say – but the latter still doesn’t.

Incidentally, we think that this sort of shift between two sorts of contrastive explanations explain why some people have denied that there are contrastive explanations of indeterministic events – explanations of why A rather than B happened in situation S. The arguments adduced to support this denial all involve comparing two cases where the histories, *H* and *H'*, leading up to the two contrasting events are intrinsically identical (see (Salmon 1984: 110; Lewis 1986: 230-1)), thus raising the question why A happened in *H* but B in *H'*. But this is

The standard libertarian response to the problem of contrastive luck is to insist that our judgments of responsibility should be based on what actually causes the action, not on comparisons with possible alternatives or on contrastive explanations (Kane 1999: 110-14; Clarke 2005: 416-19). The question, though, is *why*? Clarke (2005: 418) suggests that an argument from contrastive luck would not be successful in court, or with non-philosophers, but Mele (2006: 71-2) rightly rejects the idea that common sense should be the judge of sophisticated philosophical arguments. At least we need a *reason* to put more trust in common sense. Clarke (2005: 414-16) also suggests that making responsibility hostage to the availability of certain kinds of explanations inappropriately introduces the *pragmatics* of explanation into a question that concerns the *metaphysical grounds* for freedom and responsibility. But given the Explanation Hypothesis, this reaction is fundamentally mistaken: our thoughts about responsibility are *essentially* explanatory and pragmatic and make little sense if such features are ignored.

Obviously, the very point of Mele's argument is to highlight the element of luck involved. However, if the correctness of everyday explanatory judgments in general and judgments of responsibility in particular is relative to an explanatory background against which the judgment is made, highlighting the element of luck by asking for contrastive explanations is subtly, but essentially, changing the question whether the agent was responsible for his deed. This would mean that the judgment made after considering Mele's argument cannot really contradict our initial judgments. This parallels the conclusion in the previous section, and it raises, again, the question of which explanatory perspective we *should* take when making our judgments of moral responsibility. In the next section, we will say a few tentative words in defense of perspectives that come naturally to people when they are not considering philosophical arguments.

a different question than the question of why A rather than B happened in S. To answer the latter, what we need, roughly, is some feature of S that makes A more likely than B. (See Hitchcock 1999) for a defense of the latter claim.) To answer the former, by contrast, we need some difference between H and H' that makes A more likely in the former and B more likely in the latter.

Consequences of the Explanation Hypothesis

Thus far, we have done three things. First, taking into account the central psychological and social role of judgments to the effect P is responsible for E, we have hypothesized that such judgments keep track of whether E is explained in normal ways by some motivational structure of P that is of a kind that can be modified by reactive attitudes. Second, we have argued that this hypothesis – the Explanation Hypothesis – accounts for the force and limits of a number of everyday sources of diminished responsibility. Finally, we have argued that the Explanation Hypothesis explains the dynamics of some central aspects in the philosophical debate about the possibility of moral responsibility. All in all, this motivates looking more closely at its implications, both for further investigations in moral psychology and for the philosophical debate about the conditions for moral responsibility.²³

One particularly striking consequence of the Explanation Hypothesis is that participants of the philosophical debate about moral responsibility seem to misunderstand what the debate is about. It now seems that in putting forth arguments, they are not so much discussing a given issue as they are *changing the question* by changing the way judgments of moral responsibility are framed. In general, one should be wary of attributing such mistakes, but there are three reasons that they could be expected in this case.

The first reason is that shifts in explanatory perspectives or frames tend to be imperceptible. Few people even have concepts of such perspectives, and for good reason: an explanatory perspective is largely constituted by what is part of an explanatory background, and explanatory backgrounds are, per definition, comparatively unremarkable. In this regard, the context-relativity of explanatory claims is very different from the context-relativity of paradigmatic context-dependent expressions like pronouns (“I”, “she”, “they”), demonstratives (“that”, “this”) and temporal adverbs (“now”, “before”), where we readily and separately keep track of individuals and times that might be the *relevant* referents. The second reason is that, as we noted in our brief discussion of bystander responsibility, different explanatory perspectives are psychologically mutually exclusive; to decide which perspective

²³ To test the Explanation Hypothesis further, we are currently trying to extend its application in two directions. Firstly, we are looking at recently published and forthcoming data on intuitions concerning moral responsibility. Secondly, we are trying to argue that it provides a unified account of the many different components of responsibility that have been identified by compatibilist accounts.

to take will therefore feel very much like deciding the question at hand, especially when these perspectives are introduced by arguments rather than by stipulation or instructions to see things in a certain way. Finally, holding fixed the connection between judgments of responsibility and our willingness to take various reactive attitudes, clashing choices of explanatory perspective have conflicting practical implications. Given all this, it is only natural that there will seem to be a real conflict between judgments of responsibility between people who take different explanatory perspectives.

Although it seems permissible to attribute this kind of error, the Explanation Hypothesis doesn't *dictate* such an interpretation. If it isn't understood as a definition of moral responsibility but rather as a hypothesis about the basic criteria we apply in making these judgments, it allows that judgments made from one explanatory perspective are correct, at the exclusion of judgments made from another. If that is the case, the philosophical debate might well concern a real issue. What isn't obvious is how one should go about arguing for the exclusive correctness of judgments made from one perspective rather than another. Invoking more arguments of the sort that we have looked at in the sections on heteronomy and luck wouldn't seem to help; what such arguments would do, if successful, is to change or determine explanatory perspectives, not to show that some perspective is the right one. We need a different sort of evidence.

One suggestion would be to start from an account of when it is appropriate to hold people responsible, or to direct reactive attitudes towards them, and then understand moral responsibility to be whatever is required for people to be the appropriate targets of these attitudes or of being held responsible (Wallace 1994, e.g.). The problem is that such an account cannot build directly on our intuitions about whether people deserve blame or punishment or ought to be held responsible, for these intuitions are very closely tied to intuitions about moral responsibility, and equally subject to arguments and counterargument concerning heteronomy and luck.

Independently of whether we think that the correctness of judgments of responsibility is relative to explanatory frame or not, what seems to be needed, then, is an account of moral responsibility or of when people are the appropriate objects of reactive attitudes, that *isn't* grounded in intuitions that are directly concerned with that matter.

Such an account might well be possible. The most obvious candidate might be a rule-consequentialist account, defining moral responsibility in terms of practices of holding people responsible that have good consequences (where *the value of consequences* is independent of

considerations of responsibility). A related but different account would fit into a eudaimonia-based virtue ethics, and would define responsibility by the criteria for holding people responsible that we need to internalize to flourish under normal circumstances (where *flourishing* is understood without reference to responsibility). Further possibilities are contractualist or rationalist, perhaps asking for a maxim for attributing responsibility that we could will to be law, or one that couldn't reasonably be rejected by others (where neither *will* nor *reasonable rejection* is constrained by considerations of moral responsibility).

Obviously, we cannot assess or compare the merit of these or other possible accounts here. However, there is good reason to think that *any* plausible account will favor explanatory perspectives that are easily available for most people and that lead us to hold people responsible for most of their actions as well as for many events that depend on those actions. After all, holding and being held responsible is an integral part of many valuable human activities, and one that encourages beneficial behavior; trying to radically change this practice would therefore likely be detrimental, if radical change is at all possible. Moreover, the only *general* reasons that have been given for abandoning this practice are exactly reasons of heteronomy and luck. Since the viability of *those* reasons is exactly what is at issue, they cannot be relied upon here.²⁴ The upshot seems to be that, independently of commitment to any determinate normative approach, we have a general reason to steer our reactive attitudes and actions by judgments of responsibility made from an everyday perspective, and no general reason not to.

If this conclusion is plausible, the Explanation Hypothesis has wide-ranging implications not only for how we should understand the debate about moral responsibility, but also for the outcome of that debate; we no longer have any reason for taking skepticism about moral responsibility seriously. Even if that implication is resisted, however, we think that the explanations of a variety of judgments of moral responsibility that are offered by the Explanation Hypothesis makes it worthy of further study.

²⁴ Our conclusion here is very much in accord with Peter Strawson's (1974) argument. Strawson thought, first, that it is entirely unrealistic that we should abandon the reactive attitudes in general, and, second, that if we had a choice, it would not be settled with reference to determinism. What our analysis of central arguments in the philosophical debate has added is a further reason to dismiss even the theoretical force of the skeptical case against moral responsibility.

References

- Alicke, M. D. (1992) Culpable Causation. *Journal of Personality and Social Psychology* 63:3, pp. 368-78
- Chisholm, R. (2003) Human Freedom and the Self. In *Free Will*, ed. Milada Broukal and Gary Watson, Oxford U. P., pp. 26-38
- Clarke, R. (2005). Agent Causation and the Problem of Luck. *Pacific Philosophical Quarterly* 86, pp. 408–421
- Dennett, D. (1984) *Elbow Room: The Varieties of Free Will Worth Wanting*. Oxford U. P.
- Dennett, D. (2003) *Freedom Evolves*. Penguin
- Ekstrom, L. W. (2000) *Free Will: A Philosophical Study*. Westview Press
- Fischer, J. and Ravizza, M. (1998) *Responsibility and Control*. Cambridge U. P.
- Frankfurt, H. (1969) Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy*, 66:23, pp. 829-839
- Frankfurt, H. (1971) Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 68:1, pp. 5-20
- Hart, H. L. A. and Honoré, T. (1985) *Causation in the Law*. Oxford U. P.
- Hawthorne, J (2001) Freedom in Context. *Philosophical Studies* 104:1, pp. 63-79
- Hitchcock, C. (1999) Contrastive Explanation and the Demons of Determinism. *British Journal of the Philosophy of Science* 50, pp. 585–612
- Kane, R. (1996) *The Significance of Free Will*. Oxford: Oxford U P
- Kane, R. (1999) On Free Will, Responsibility and Indeterminism: Responses to Clarke, Haji, and Mele. *Philosophical Explorations*, 2:2, pp. 105-121
- Knobe, J., Fraser, B. (2008). Causal Judgment and Moral Judgment: Two Experiments. In *Moral Psychology* Vol 2, ed. Sinnott-Armstrong, W., MIT Press, pp. 441-47
- Lewis, D. K. (1986a) Causal Explanation. In *Philosophical Papers, Vol II*, Oxford U. P., pp. 214-40
- Mele, A. (1995) *Autonomous agents –From self control to autonomy*. New York: Oxford U P
- Mele, A. (2005) Libertarianism, Luck, and Control. *Pacific Philosophical Quarterly* 86, pp. 381–407
- Mele, A. (2006) *Free Will and Luck*. New York: Oxford U P
- Nagel, T. (1979) *Moral Luck in Mortal Questions*. Cambridge: Cambridge U P
- Nahmias, E; Coates, J; Kvaran. T. (2007) Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions. *Midwest studies in Philosophy* XXXI (2007), pp. 214-42
- Nichols, S; Knobe, J. (2007) Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions, *Noûs* 41:4, pp. 663-685
- Northcott, R. (2008) Causation and Contrast Classes, *Philosophical Studies* 139:1, pp. 111-23
- O'Connor, T. (2000) *Persons and Causes: The Metaphysics of Free Will*. Oxford U. P.

- Pereboom, D. (2001) *Living without Free Will*. Cambridge U. P.
- Persson (2005) *The Retreat of Reason*. Oxford U. P.
- Rieber, S. (2006) Free Will and Contextualism. *Philosophical Studies* 129:2, pp. 223-52
- Salmon, W. C. (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton U. P.
- Sher, G (2006) Out of control. *Ethics* 116, pp. 285-301
- Smilansky, S. (2000) *Free Will and Illusion*. Oxford: Oxford U P
- Smilansky, S. (2003) Compatibilism: The Argument from Shallowness. *Philosophical studies* 115, pp. 257-282.
- Smith, A. (2008) On Being Responsible and Holding Responsible. *The Journal of Ethics* 11, pp. 465-84
- Strawson, P. (1974) Freedom and Resentment. In *Freedom and Resentment and Other Essays*. Methuen & Co., pp. 1-25
- Strawson, G. (1986) *Freedom and Belief*. Oxford U. P.
- Strawson, G. (1994) The Impossibility of Moral Responsibility. *Philosophical Studies* 75, pp. 5-24
- Van Inwagen, P. (1983) *An Essay on Free Will*. Clarendon Press
- van Inwagen, P. (2000) Free Will Remains a Mystery. *Philosophical Perspectives* 14, pp. 1–19
- Wallace, R. J. (1994) *Responsibility and the Moral Sentiments*. Harvard U. P.
- Watson, G. (1996) *Philosophical Topics*, 24:2, pp. 227-48.
- Wolf, S. (1990) *Freedom within Reason*. Oxford U. P.
- Woolfolk, R. L; Doris, J. M.; Darley, J. M. (2006) Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility. *Cognition* 100, pp. 281-301