The Explanatory Component of Moral Responsibility

GUNNAR BJÖRNSSON
Department of Culture and Communication, Linköping University
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg

KARL PERSSON
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg

Introduction

People who have thought long and hard about the requisites for moral responsibility are still in deep disagreement. While some feel strongly that determination of choices and actions by causes outside the agent's control undermines responsibility,[i] 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.[ii] And many disagree about whether luck of various kinds is compatible with moral responsibility and to what extent responsible actions must be fully determined by rational deliberation.[iii]

As we will see, some of the most important arguments supplied in these controversies are effective insofar as they lead us to *focus* on certain aspects 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.[iv]

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? Clearly, more arguments are needed; what is far from clear is what sort of argument we should be looking for.

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 psychological hypothesis about judgments of moral responsibility, a hypothesis according to which such judgments are a species of *explanatory* judgments.

Second, we will show how this model can account not only for factors that affect the degrees to which we assign moral responsibility in ordinary life, but also for 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 will suggest that the perspective relativity of responsibility judgments has important methodological consequences for the debate about moral responsibility. Ultimately, it provides support for judgments of responsibility that rely on everyday perspectives while undermining those that rely on perspectives induced by skeptical philosophical arguments.

The three components of moral responsibility

There are two reasons to take the psychological hypothesis that we will propose here seriously. The most important reason, ultimately, and the one that we will focus on throughout most of this paper, is that the hypothesis explains both everyday responsibility judgments and judgments made in response to philosophical arguments. The other reason is that we can expect something like it to be correct given the role that judgments of moral responsibility play in our lives. The latter, etiological, reason is more speculative, but since it also serves to introduce the hypothesis, it merits a sketch.

It is well known that judgments of responsibility govern moral reactive attitudes and behaviors: guilt, resentment, blaming, punishing and demanding compensation, as well as moral admiration and moral praise. Most people think that someone deserves praise, blame or punishment for something only to the extent that she is responsible for it. Similarly, the degree to which a person in need is seen as responsible for her predicament often affects the degree to which she is seen as entitled to help, and the degree to which she is seen as responsible for something good affects the degree to which she has a right to corresponding rewards. In a slogan: judgments to the effect that an agent *is* responsible for something affect the extent to which it is seen as appropriate to *hold* that agent responsible for it.[v]

It seems undeniable that most of our concern with moral responsibility stems from this connection between *taking* people to be responsible for something and *holding* them

responsible. This is not to say that taking someone to be responsible for some wrongdoing is *always* accompanied by having certain attitudes, or by a disposition to let her suffer or enjoy the consequences of her decisions. Various psychological circumstances might promote quite different attitudes, and further reasons might dictate different actions in particular cases.[vi] But reactive attitudes typically do abate or disappear when we realize that their object was not responsible for the event or action that had triggered them, and someone's being responsible for something seems to make it at least prima facie permissible to hold her responsible for it. Furthermore, in saying that concern with moral responsibility stems mostly from concern with whether to hold people responsible, we do not rule out that one can be seen as responsible for things that are neither praise- or blameworthy, nor good or bad. We might be fully responsible for our exact choice of route as we stroll through a park, or for a trivial pattern made with a stick in the sand. But talk about responsibility in general and moral responsibility in particular is most natural in contexts where what someone is responsible for is something of positive or negative importance.

If peoples' concern with moral responsibility is mainly driven by a concern with *whom to hold responsible for what*, it is natural to think that the ordinary concept of moral responsibility has been shaped by conscious and unconscious interests and concerns that govern our practice of holding people responsible. Largely, such concerns seem to be directed at controlling and shaping motivational structures—priorities, values, preferences, desires, behavioral and emotional habits, etc.—in order to promote and prevent certain kinds of behaviors and outcomes. For example, indignation over some action is normally placated by the agent's expression of motivation to avoid such actions in the future, and particular practices of holding people responsible are frequently motivated by the idea that this will produce more "responsible" behavior, or that their absence would encourage childishness and irresponsibility.

In order for practices of holding people responsible to properly control and shape motivational structures and promote and prevent various kinds of behaviors or other events, they need to be directed towards the sort of *motivational structures that* (a) *explain these events in systematic ways* and (b) *respond in the appropriate way to the agent's being held responsible*. Moreover, since actual experience provides much more powerful displays of both the presence of the relevant motivational structures and the explanatory connection between these structures and the relevant events, we need to hold people responsible on

occasions when these motivational structures (c) *are significant parts of the explanation of such events*.

The suggestion, then, is that whatever mechanisms determine whom we hold responsible for what will tend to direct our attitudes towards motivational structures that satisfy (a), (b) and (c). But we have also argued that the everyday concept of moral responsibility is a significant part of those mechanisms, having as its main psychological function to direct practices of holding people responsible. Our hypothesis, therefore, is that people take P to be morally responsible for E when they take some motivational structure of P to satisfy (a), (b) and (c). Call this the *Explanation Hypothesis.*

A number of clarifying remarks and caveats are needed before we look closer at the three conditions. First, although we take considerations of moral influence to have shaped the everyday concept of moral responsibility and its connection to practices of holding people responsible, we are not thereby saying that such considerations *justify* the practice of holding people responsible. The Explanation Hypothesis concerns how people actually make judgments of moral responsibility, not how such judgments ought to be made.

Second, although we take considerations of moral influence to have shaped the everyday concept of moral responsibility, the Explanation Hypothesis implies that *applications* of that concept to an agent and an outcome are completely insensitive to the effects of actually holding the agent responsible for that outcome. Such applications only look at whether motivation of some relevant type is part of a significant explanation of the outcome (more on this later).[vii]

Third, the Explanation Hypothesis is a psychological hypothesis rather than a semantic analysis or a metaphysical claim. It tells us what criteria people use in determining whether someone is responsible for something, but it does not say whether these criteria define the essence of moral responsibility.

Fourth, one might worry that the various ways of holding people responsible display very little unity, and that this undermines the explanatory story; being indignant, rewarding someone for diligent service and letting starving artists starve might seem to have very little in common. But the commonality needed for our etiological explanation is only (i) that these various ways of holding people responsible are governed in large part by our judgments of moral responsibility, (ii) that they normally need for the performance of their social function that (a), (b) and (c) hold, and (iii) that our interest in making judgments of responsibility is explained, in part, by our interest in whether to hold people responsible in that way.

Fifth, 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 a purely causal or explanatory sense of responsibility according to which the question "What is responsible for E?" is interchangeable with "Why did E happen?", as well as an institutional or social sense 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").[viii] 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 condition (c).[ix] Nevertheless, we suggest that most of our strongest intuitions about moral responsibility, and the intuitions that are very much driving the philosophical debate, are tied to (a), (b) and (c).

With these remarks in mind, here is a more complete formulation of the Explanation Hypothesis:

> *Explanation Hypothesis:* People take P to be morally responsible for E to the extent that they take E to be an outcome of a type O and take P to have a motivational structure S of type M such that GET, RR and ER hold:
>
> > *General Explanatory Tendency (GET)*: Motivational structures of type M are significant parts of a reasonably common sort of explanation of outcomes of type O.
> >
> > *Reactive Response-ability (RR):* Motivational structures of type M tend to respond in the right way to agents being held responsible for realizing or not preventing outcomes of type O.
> >
> > *Explanatory Responsibility (ER)*: The case in question instantiates the right sort of general explanatory tendency: S is part of a significant explanation of E of the sort mentioned in GET.[x]

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 every event that happens to be explained by our motivational states. For example, if Mr. Black starts stalking Mr. Green because Mr. Green likes to wear plaid vests and because Mr. Black is obsessed with people who like plaid vests, Mr. Green is not thereby responsible for being stalked by Mr. Black. Since it is not *generally* the case that liking something explains being stalked, this case violates GET: we think of the sort of condition on which the explanation relies as being highly unusual. The explanatory relations that most clearly *do* satisfy GET involve explanations of outcomes with reference to preferences that these 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 concern to ensure such outcomes, or by their lack of sufficient concern to avoid such outcomes.[xi] Ceteris paribus, we will take George to be responsible for missing a deadline if we think that he missed it for lack of diligence; we will take him to be responsible for his daughter's improvement in school if we think that it happened because he really wanted her to succeed and spent a lot of time helping her do her homework.

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 in a crash, we might take him to be responsible for the event 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 the accident as responsive in the right way. By contrast, RR seems to be undermined in the case of overwhelming motivational states, when the agent has a general incapacity to respond with self-directed reactive moral attitudes such as guilt and shame, or when she lacks self-control in general. In conjunction with ER, RR thus 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 accounts that take moral responsibility to be compatible with determinism.

To understand RR, it is also important to see that it can be satisfied to degrees. Thus, the motivational structures of a dog or a child might be responsive to some but not all of the ways we have of holding people responsible. In such cases, we can expect people to make

uncertain, conflicted or qualified attributions of moral responsibility and desert and withhold reactions that demand a more sophisticated response than, say, simple conditioning.[xii]

Finally, notice that RR leaves open how people individuate types of motivation and thus whether a token motivational structure will be understood as an instance of a type that satisfies RR. Different ways of typing motivational structures is a possible source of differences in responsibility judgments.

Enough has been said to provide a rough idea about the content of GET and RR. In what follows, our primary concern will be 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 would typically not be held responsible for missing a meeting if I got stuck between floors in a failing elevator on the way there. In one form or other, requirements like these are also part and parcel of most compatibilist accounts.

ER has much more interesting consequences, however, having to do with the fact that it restricts our attribution of responsibility to cases where the motivational structure in question is part of a *significant* explanation of the event. We will explain this notion, and show how this has important consequences for our everyday notion of moral responsibility.

Significant explanations 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 condition that are especially interesting or relevant given our explanatory interests. When ER demands that the agent's motivational structure should be part of a *significant* explanation of the target action or event, this is what is meant.

Typically, a significant explanans, X, only provides an explanation of why E took place given a number of further conditions, C. 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 we take X to be a significant aspect of what explains E is whether X is more 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.[xiii] Consider the force and limits of six kinds of everyday excuses, or considerations that are generally taken to lower moral responsibility:

1. He was forced to do it. The Explanation Hypothesis explains why various degrees of external force, threat and constraint are taken to 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 perceived 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.[xiv] 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 on 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 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 seems to be a significant reduction of the clerk's responsibility for the bank's loss only in the second case.

In the first of these cases, unlike in cases of overwhelming physical force or obstacles, the agent's motivational structure is never *bypassed*, but is part of a complete explanation of the outcome: had the clerk been more concerned not to give in to threats than to save lives, the outcome would have been different. It is just that it is understood as part of the explanatory background rather than seen as a *significant* explanans.

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 and consider what to do *given the circumstances that include the threat*. From this perspective, the clerk's motivational structure is the variable determining the choice.[xv] 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 does not 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 she crashes into the braking car in front. Knowing this, we do not 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 a 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 the first case, the receptionist is merely doing his job; in the second he deliberately violates strict orders not to let unauthorized people into the building. In both cases, the child would have died had the receptionist not unlocked the door. When we want to explain *why* the child was saved, we are likely to mention the receptionist's willingness to put the child's life before the rules in the second case, but unlikely to mention his motivation in the first where it merely conforms to baseline normative expectations. 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 apartments are burglarized. In one case, the receptionist finds the people that are let in a little bit suspicious, 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 higher degree of responsibility in the second case, where the receptionist's willingness to unlock the door in spite of the rules would be naturally taken to explain why the apartments got burglarized.[xvi]

4. 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 one hoodlum, Hank, happens to witness when another hoodlum, Hogan, yanks the purse from an old lady's hand and takes off. Hank could have stopped Hogan, but was more interested in how he might emulate Hogan's technique. If asked why the lady no longer has her purse, we will naturally focus first on Hogan's action, and only later, it at all, on Hank's inaction. Consequently, we will primarily assign responsibility for the lady's loss to Hogan, even though Hank could have made a difference.

Being a passive witness rather than actively pursuing an outcome does not always remove moral responsibility completely, and ER explains that too. According to ER, the bystander's inaction could make her responsible if it were *remarkable* that she lacked the motivation to intervene. Suppose that Linda, a police officer, saw Hogan's deed, and suppose that she could easily have stopped the robbery but decided not to 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 about protecting the public. And just as ER predicts, Linda now seems to be morally responsible for the lady's loss.

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 does not 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.

5. 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. Compare our previous case where Linda passively watched Hogan steal the lady's purse with a second case in which the Hogan acts behind Linda's back and where she could not hear what was happening. In the first case it is natural to explain the hoodlum's successful robbery with reference to Linda's motivational structure, but not in the second. And, as ER predicts, we are now unwilling to assign moral responsibility to Linda for the success of the robbery.

ER also predicts that moral responsibility sometimes survives ignorance. Suppose that Linda failed to notice the robbery because she was busy listening to the dog racing results on the radio while on patrol duty. Now her ignorance seems to be the result of a remarkable disregard for her job: *Why was the robbery successful? Because Linda found the dog racing results more important than the street life!* As ER predicts, Linda is now found at least somewhat morally responsible for the success of the robbery.

6. It wasn't her initiative. Finally, ER neatly explains why someone who is actively pursuing an end and engages others in the effort tends to be seen as more responsible for achieving it than those who 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 concern and 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 violate or exceed normative expectations. However, the evidence for the Explanation Hypothesis does not 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, relative to our explanatory interests, of a complete normal explanation of that event. 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 heteronomy; in the next, we 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.[xvii]

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 reference to the agent's motivational structure strikes us as providing significant explanations of these actions, and so, given ER, changes our intuitive 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 know the following:

> 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 could not 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 pick 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 by cars 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 is not 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 a straightforward explanation (road work, parade, accident) 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.

Notice that our claim is not that peoples' explanatory judgments are the upshot of conscious reasoning that invokes the considerations of explanatory power and economy and concludes that only (d) provides a *good* explanation of the late arrival. The idea, rather, is that these considerations affect what strikes people as a *significant* explanans, one that stands out among other conditions that are nevertheless assumed to be parts of a *full* causal explanation of the outcome also involving (a), (b), (c), (e) and (f).

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 has not

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's 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 would not expect straightforward explanations of why someone cares very little about his dog, or likes Beethoven (or at least none that does not invoke some other RR and GET-satisfying motivational structure that itself lacks such an explanation).[xviii]

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 significant explanations of 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 are often parts of significant explanations of particular actions or events. 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 "indeterministic factors". This removes perceived explanatory significance of motivational structures in two steps.

To begin with, it eradicates differences in perceived complexity between, on the one 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* summarizes what would otherwise have to be understood as enormously complex sets of prior conditions and explanatory relations. This means that there is no 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) rather than by (d). 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).[xix]

The upshot is an explanatory regress of the kind that pushed us from (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.[xx] Given ER, that means that agents will not seem responsible for their actions, even though their motivational states are still understood to be part of a full causal explanation of these actions.

The last point is important: we are not saying that Strawson or others who are impressed by regress arguments of this sort will deny that most of our actions are explained by our motivational states; it is just that while contemplating the argument, such explanations will not strike them as significant. Also, it is not to say that *everyone* will be impressed by regress arguments. There might be considerable individual variation, and variation from case to case, depending on how inclined people are to take the abstract perspective and how wedded they are to more concrete perspectives. For example, 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.[xxi]

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 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.[xxii] 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 that lack 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 judgments 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.

This would be 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. In effect, the same question was raised by some of the everyday excuses considered in the previous section. For example, we noticed that whether one sees the clerk who handed over the bank's money to robbers as responsible for the bank's loss depends on whether one takes the robbers' threat or the clerk's motivation as part of the explanatory background. Since both are understood to be part of a *full* explanation of the bank's loss, the background relativity of explanatory judgments would make it possible both to deny and to affirm that the clerk is responsible for the bank's loss. The question would be *what perspective to take.* (By contrast, the presence of overwhelming external force or the absence of knowledge might rob a certain motivational structure of *all* explanatory relevance for an outcome, whether it is seen as part of the background or not. To stress that someone was physically prevented from doing something is not to move the assumption of a causal connection between motivational structure and characteristic outcome from explanatory background to foreground, but to deny that there was any such connection.)

As we will see in the next section, the question of what perspective to take in making explanatory judgments is also raised by another family of skeptical arguments, and we will begin to address the question before closing, arguing 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 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 or stable motivational structures. Relative to our stable 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.

It is clear enough that human action is subject to luck of this sort. What we do is not always completely determined by our stable rational thought or stable motivational patterns of the sort that is affected by the practice of holding people responsible. Sometimes our decision processes are influenced by external factors in non-systematic ways, grabbing our attention and prompting decisions that would have made little sense a moment earlier and made little sense in retrospect; at other times, reason and preferences seem to leave a question undecided until, for no clear reason, we just decide one way rather than another.

At first glance, cases like these might be compatible with moral responsibility, as we might nevertheless have a high enough degree of rational control *most* of the time, and since moral responsibility seems compatible with *some* lack of rational control. 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 stable motivational structure just before the decision. Given his wants, desires, habits, and rational evaluation of behavior, his decision *could* have gone both ways. This could either be because the causal connection between motivational structure, rational thought 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.

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. John's motivational structure 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:[xxiii] they show that indeterminism need not seriously undermine responsibility. However, Al Mele (2005; 2006, ch. 3) has recently put the problem of luck in a new and intuitively forceful way. Consider two worlds: the actual world, W, in which John decides to kill Bill, and a possible world, W', which is indiscernible from the actual world up to the moment of John's decision but where John does not 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 takes responsibility to rest on cases like John's remain in an uneasy position.

Libertarians are not alone in that uneasy position, however. Even in a deterministic universe, a variety of events that the agent cannot predict and over which she lacks rational control might influence her decisions: chemical events in her neurology, say, or external events that affect her senses and happen to direct her attention in certain ways rather than others. Mele's challenge thus 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.

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 could 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. We 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, thus changing the explanatory frame that was involved in coming to the prior conclusion that John was responsible for killing Bill. 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 structures—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 particularly explanatory.[xxiv]

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 support of perspectives that come naturally to people when they are not considering philosophical arguments.

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 that P is responsible for E, we have hypothesized that such judgments keep track of whether E is significantly explained in normal ways by some motivational structure of P that is of a kind that can be modified by holding people responsible for effects of that structure. 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 arguments in the philosophical debate about the possibility of moral responsibility.[xxv] All in all, this motivates looking more closely at its implications, both for further investigations in moral psychology and for the philosophical debate about moral responsibility.

If the Explanation Hypothesis is correct, responsibility judgments depend on explanatory perspectives, as we have illustrated by a number of cases. This perspective relativity can be understood in different ways, depending on one's preferred semantics for responsibility judgments. One option is to take the criteria that we actually use in judging whether someone

is responsible for something as defining the truth-conditions of the resulting judgments. Judgments to the effect that P is responsible for E would thus be true if and only if GET, RR and ER hold. If substantial explanatory claims are true or false relative to explanatory perspectives, the same would hold for responsibility judgments, giving us a contextualist analysis of moral responsibility.[xxvi] A second option is realist—cognitivist but anti-contextualist—and takes there to be a perspective-independent truth about whether someone is responsible for something, thus conforming to the standard assumption that it is an objective matter of fact whether responsibility is compatible with determinism or forms of luck. On this option, the perspective relativity is epistemic or doxastic, not semantic. A third option, sometimes associated with Peter Strawson's "Freedom and Resentment" (1974), follows the lead of expressivist theories in ethics and rejects analyses in terms of substantial truth-conditions, instead analyzing responsibility judgments as expressions of (defeasible) dispositions to hold people responsible. On an expressivist accommodation of the Explanation Hypothesis, explanatory perspectives are among the factors that affect these dispositions.

These three semantic interpretations of responsibility judgments all have complex strengths and weaknesses, so this is not the place to decide between them. But regardless of which option one chooses, there seem to be reasons to accept and be guided by responsibility judgments made from everyday perspectives and to reject judgments made from the perspectives introduced by skeptical philosophical arguments.

Suppose first that the best semantics for responsibility judgments is a form of realism. On this view, there is a context-independent fact of the matter of what moral responsibility consists in; the difficulty lies in finding out what that fact is. Given the Explanation Hypothesis, further arguments of the sort we looked at in the sections on heteronomy and luck are likely to be just more ways to affect explanatory perspectives, rather than to show that the verdicts made from one such perspective is correct. We would need a different kind of argument.

One possibility would be to appeal to some substantial theory of reference and try to determine what responsibility is by trying to determine what it is that stands in the reference relation to our concept of responsibility. In order not to yield a form of contextualism, such a theory must forsake heavy reliance on our intuitions about responsibility, since these intuitions vary with explanatory perspective. The trick would be to capture the relation that we are interested in when we are interested in moral responsibility, without relying on our criteria for determining whether someone is responsible for something.

Some theories of reference would ask us to identify the relation of someone's being responsible for something with whatever relation best fits our platitudes about moral responsibility, say, or with whatever relation we can be seen as having accumulated knowledge about when using our concept of moral responsibility.[xxvii] But it is not clear how that would help us. The problem with the first of these particular suggestions is that it is unclear how to determine what the platitudes are or what best fits them given the perspective dependence of our responsibility judgments. The problem with the second is that it is part of the issue at hand whether we are accumulating *knowledge* about responsibility under everyday explanatory perspectives.

The most promising venue, we think, would be to look at our practice of holding people responsible *as a whole* and look at what *it* seems to be concerned with. Once we take that perspective, the concerns with influencing motivation by holding people responsible that motivated the Explanation Hypothesis are again highly relevant; it is plausible that a concept functioning according to the Explanation Hypothesis would have been shaped by just such concerns. If that is correct, it also seems plausible to say that what our concept of moral responsibility is *meant* to capture is what GET, RR and ER do capture when cases are assessed from stable everyday perspectives involving no factual errors. That is what our judgments have needed to keep track of in order to perform what we might call their "etiological" function.[xxviii] Admittedly, this suggestion is rather vague and might be resisted for a variety of reasons, but it is not clear that any other appeal to substantial notions of reference will be of *more* help for the anti-contextualist cognitivist.

Another possibility is to invoke the connection between responsibility judgments and various normative judgments. In particular, it is tempting to invoke the fact that it seems wrong, ceteris paribus, to hold an agent responsible who is not responsible. Moral responsibility would then be whatever prerequisite for holding someone responsible that our concept of responsibility is best seen as tracking. To determine what it is to be morally responsible for something and when people are morally responsible, we should thus turn to the question of when it is appropriate to hold them responsible (cf. Wallace 1994, e.g.).

This question, it seems, is also the very question that needs to be asked given contextualist or expressivist interpretations of responsibility judgments and the Explanation Hypothesis. On a contextualist understanding, there are many forms of moral responsibility, picked out from different explanatory perspectives. To determine whether any one of them identifies a precondition for holding someone responsible, we need to have a prior idea about

when holding someone responsible is appropriate. On an expressivist understanding, the question of whether an agent is morally responsible just *is* the question of whether to be (defeasibly) disposed to hold that agent responsible, and it is hard to see how that would be decided without taking a stance on the normative preconditions for holding someone responsible.

Regardless of semantics, then, the *important* question about moral responsibility might well turn out to be the question of what the normative preconditions are for holding someone responsible (assuming that realist appeals to substantial theories of reference fail). The problem we are facing 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 counterarguments concerning heteronomy and luck. What seems to be needed is an account of moral responsibility or of when people are the appropriate objects of reactive attitudes that *is not* based on intuitions that are directly concerned with that matter.

Such an account might well be possible. The most obvious candidate might be a consequentialist account, defining moral responsibility in terms of practices of holding people responsible that have good consequences (where *the value of consequences* is understood independently 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 could not reasonably be rejected by others (where neither *the will* nor *reasonable rejection* is antecedently 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, *pace* (Strawson 1974)). 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 what is at issue, they cannot be relied upon here. The upshot seems to be that, independently of commitment to any determinate normative approach, and independently of commitment to contextualism, realism or expressivism, 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 (cf. Vargas 2007, 154-60).

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. Whether or not we should take skeptical arguments to change the question when they change our explanatory frames, we no longer have any reason for taking skepticism about moral responsibility seriously. Even if that implication is resisted, however, the explanations of a variety of judgments of moral responsibility offered here are enough to make the Explanation Hypothesis worthy of further study.[xxix]

---

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

[ii] Frankfurt (1969; 1971), Wallace (1994), Persson (2005)

[iii] 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 and Ravizza (1998) think that those problems are surmountable.

[iv] For effects on "folk" intuitions, see (Nichols and Knobe 2007) and (Nahmias, Coates and Kvaran 2007), e.g.. For discussion, see (Björnsson and Persson 2009; *manuscript A*).

[v] The locus classicus on the importance of the connection between reactive attitudes and moral responsibility is Sir Peter Strawson's seminal (1974) paper "Freedom and Resentment".

It is worth noting that these practices of holding people responsible that we discuss here might belong specifically to western culture (Sommers 2009).

[vi] Angela Smith (2007) stresses that moral responsibility is 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. Manuel Vargas (2008, 93-94) points out that it would be possible, without conceptual confusion, to have as a policy not to hold people responsible for the first impolite remark they are responsible for, and that this policy could make it inappropriate or unfair to hold a first-time offenders responsible. See also (Scanlon 2008).

[vii] In these ways, the Explanation Hypothesis is importantly unlike traditional moral influence theories of responsibility. For an overview of problems with such theories, see (Vargas 2008).

[viii] For discussion of relations between some notions of responsibility, see (Watson 1996).

[ix] 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.

[x] 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. For discussion of how this account applies to joint or collective responsibility, see Björnsson (forthcoming). 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.

[xi] Typically, these explanations demand that the agent *knows* about a possible outcome and about available means to prevent or promote it; such conditions provide the relevant C for the most common explanation of outcomes in terms of motivational structures. 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 need 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).

[xii] Ingmar Persson is one of many who contrasts full responsibility with mere susceptibility to manipulation, and takes the former to imply that the agent is "conceptually equipped to view the [Reward–Punishment] practice as being applied to herself because she has caused something good or bad for some being" (Persson 2005, 404-6). This condition captures a form of reactive response-ability required for the normal success of "mature" ways of holding responsible.

[xiii] The way we apply this idea to judgments of moral responsibility is similar to the way Hart and Honoré (1985, 33-44) applied it to legal responsibility.

[xiv] If this is not obvious enough, it is confirmed by Woolfolk et. al. (2006), who report experiments where increasing coercion shifted assigned responsibility from agent to coercer.

[xv] 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 do not always seem completely unrealistic.

[xvi] In general, normative expectations seem to affect explanatory interests by affecting allocations of explanatory conditions to foreground and background, along with statistically based expectations. 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 and an assistant each took one of the last two pens. Later during the day, one of the assistants 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 the professor's and the assistant's physical behavior was the same. It also seems quite clear that we would be more willing to attribute responsibility for the problem to the professor than to the assistant.

Since differences in normative expectations can affect what goes into the explanatory background, they can strongly influence responsibility judgments. For example, in a society where it is taken as background that men act on sexual urges, women who trigger such urges are more likely to be seen as morally responsible for the actions of a rapist, and for being raped.

[xvii] See also (van Inwagen 1983; 2000), (Strawson 1986), (Kane 1996) and (Pereboom 2001).

[xviii] H. L. A. Hart and Tony 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.

[xix] Steven Rieber (2006) proposes a contextualist account of free action which gives much simpler explanations of why regress arguments undermine attributions of free action: they do so simply by mentioning prior causes. But mentioning prior causes is *not* enough to undermine freedom; it is common to mention that an agent does so-and-so because of gender, race, origin, income bracket and so forth in much the same breath as that agent is held responsible for these actions.

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

[xxi] 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 in (Björnsson and Persson 2009; *manuscript*). For a discussion of interpersonal variation in responsibility judgments, see (Feltz and Cokely 2009), e.g..

[xxii] 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.

[xxiii] 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 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.

[xxiv] Clarke (2005, 415) subtly misrepresents the contrastive explanation involved in Mele's argument. Mele's 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 does not.

Incidentally, we think that this sort of shift between two sorts of contrastive explanations explains 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'. Since there is no answer to that question, people have concluded that indeterminism precludes contrastive explanations. But the question that lacks an answer is 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.

[xxv] There are, of course, other skeptical arguments against moral responsibility than those considered here. For example, Derk Pereboom (2001, 112-26) contends that there is no important difference between agents who have been subject to covert manipulation and agents whose actions are determined by causes outside their control. Since the former are not responsible, neither are the latter.

Others have argued that the two cases *are* importantly different in that agents' rational capacities are bypassed only in the manipulation cases (cf. Fischer & Ravizza 1998, Fischer 2006, Mele 1995; 2006), but Pereboom (2005; 2007; 2008) has stood his ground.

In a related but importantly different argument, Al Mele (1995; 2006, e.g.) employs a case where a goddess designs a zygote in a deterministic universe in order to deliberately ensure that the person developing from the zygote will perform a certain action much later. Here, there is no plausible sense in which the person's rational capacities are *bypassed*, but such a case nevertheless tends to evoke the intuition that the person is not fully responsible for that action; the challenge for the compatibilist is to either explain away the intuition, or explain why ordinary causal determinism is importantly different from manipulation by design in some other regard.

Elsewhere (Björnsson and Persson *manuscript B*), we (i) argue that the Explanation Hypothesis can explain intuitions prompted by arguments from different manipulation, thus further strengthening the hypothesis, and (ii) offer a defense against skeptical or incompatibilist conclusions, similar to that offered in this paper, improving on prior compatibilist strategies by offering a more general diagnosis and a way of dealing equally well with overt manipulation working through an agent's rational deliberation.

[xxvi] Forms of contextualism about free action have been proposed by a number of authors (Hawthorne 2001, Dennett 2003, 93 e.g., Rieber 2006). John 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. Daniel Dennett argues that judgments made from the God's eye perspective are different from everyday judgments, 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. Although Rieber explicitly refrains from drawing conclusions about moral responsibility, an extended version of his proposal might be closest to a contextualist version of the Explanation Hypothesis. But there are differences. The latter gets more wide-ranging support from 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, and has an easier time accounting for disagreements about responsibility and for the fact that many people accept or reject compatibilism independently of conversational perspective. Moreover, as mentioned in a footnote above, Rieber's analysis seems to predict that the mere mentioning of prior causes of our actions should undermine attributions of freedom (or responsibility), which it clearly does not (see Nichols & Knobe 2007, e.g.).

Richard Feldman (2004) takes contextualism about free will to task for (i) failing to account for the sense that incompatibilism threatens ordinary ascriptions of freedom, for (ii) granting that incompatibilists are right in saying what they say (in the proper contexts) and for (iii) failing to address the real issue about free will. We agree that the hard questions remain—the normative questions about when it is right to hold people responsible—, but think that contextualists can explain the sense that incompatibilist judgments are in conflict with ordinary ascriptions of freedom (or responsibility). In general, one should be wary of interpreting people as mistaken about their own activities in the way that contextualism implies. But there are three reasons that this particular mistake could be expected. 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, by 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 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, as opposed to a mere difference in the questions that are asked.

[xxvii] Compare Richard Boyd's (1988) and Michael Smith's (1994) approaches to moral realism.

[xxviii] Compare teleosemantic approaches to content and reference (Millikan 1984, Papineau 1993, Boyd 2003a; 2003b)

[xxix] The paper has benefited from comments by audiences at the research seminars in practical philosophy at the University of Gothenburg and Stockholm University as well as at The Fifth Interuniversity Workshop on Art, Mind and Morals, Palma de Mallorca, December 2008. We are especially grateful for comments from Ingmar Persson, Göran Duus-Otterström, David Alm, Shaun Nichols, Manuel Vargas and an anonymous referee for Noûs.

References

Alicke, M. D. (1992) 'Culpable Causation,' *Journal of Personality and Social Psychology* 63:3, pp. 368-78

Björnsson, G. (2007) 'How Effects Depend on Their Causes, Why Causal Transitivity Fails, and Why We Care about Causation,' *Philosophical Studies* 133:3, pp. 349-90

Björnsson, G. (*forthcoming*) 'Joint Responsibility without Individual Control: Applying the Explanation Hypothesis,' *Compatibilist Responsibility: Beyond Free Will and Determinism*, eds. Jeroen van den Hoven, Ibo van de Poel and Nicole Vincent

Björnsson, G.; Persson, K. (2009) "Judgments of Moral Responsibility: A Unified Account", *Society for Philosophy and Psychology, 35th Annual Meeting 2009*, http://philsci-archive.pitt.edu/archive/00004633/

Björnsson, G.; Persson, K. (*manuscript A*) 'A Unified Empirical Account of Responsibility Judgments'

Björnsson, G.; Persson, K. (*manuscript B*) 'The Manipulation in Manipulation Arguments against Moral Responsibility'

Boyd, R. (1988) 'How to be a Moral Realist,' *Essays on Moral Realism*, Cornell U. P., ed. G. Sayre-McCord, Cornell U. P., pp. 181-228

Boyd, R. (2003a): 'Finite Beings, Finite Good: The Semantics, Metaphysics and Ethics of Naturalist Consequentialism, Part 1'. *Philosophy and Phenomenological Research* 66, pp. 505-53

Boyd, R. (2003b): 'Finite Beings, Finite Good: The Semantics, Metaphysics and Ethics of Naturalist Consequentialism, Part 2'. *Philosophy and Phenomenological Research* 67, pp. 24-47

Chisholm, R. (2003) 'Human Freedom and the Self,' *Free Will*, ed. M. Broukal and G. 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

Feldman, R. (2004) 'Freedom and Contextualism,' *Freedom and Determinism*, ed. J. K. Campbell, M. O'Rourke, D. Shier

Feltz, A.; Cokely, E. T. (2009) 'Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism,' *Consciousness and Cognition* 18, pp. 342-350

Fischer, J. and Ravizza, M. (1998) *Responsibility and Control*. Cambridge U. P.

Fischer, J. (2006) *My Way: Essays on Moral Responsibility*. Oxford 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 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,' *Moral Psychology* Vol 2, ed. Sinnott-Armstrong, W., MIT Press, pp. 441-47

Lewis, D. K. (1986) 'Causal Explanation,' *Philosophical Papers, Vol II*, Oxford U. P., pp. 214-40

Mele, A. (1995) *Autonomous agents: From self control to autonomy*, Oxford U P

Mele, A. (2005) 'Libertarianism, Luck, and Control,' *Pacific Philosophical Quarterly* 86, pp. 381–407

Mele. A. (2006) *Free Will and Luck*, Oxford U. P.

Millikan, R. (1984) *Language, Thought, and Other Biological Categories: New Foundations for Realism*, MIT Press

Nagel, T. (1979) 'Moral Luck,' *Mortal Questions*, 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.

Papineau, D. (1993) *Philosophical Naturalism*, Blackwell

Pereboom, D. (2001) *Living without Free Will*, Cambridge U. P.

Pereboom, D. (2005) 'Defending Hard Incompatibilism,' *Midwest Studies in Philosophy*, XXIX

Pereboom, D. (2007) *Four Views on Free Will*, Blackwell

Pereboom, D. (2008) 'A Hard-line Reply to the Multiple-Case Manipulation Argument', *Philosophy and Phenomenological Research* 77:1, pp. 160-70

Persson, I. (2005) *The Retreat of Reason*, Oxford U. P.

Rieber, S. (2006) 'Free Will and Contextualism,' *Philosophical Studies* 129:2, pp. 223-52f

Salmon, W. C. (1984) *Scientific Explanation and the Causal Structure of the World*, Princeton U. P.

Scanlon, T.M. (2008) *Moral dimensions*, Harvard U. P.

Sher, G (2006) 'Out of control,' *Ethics* 116, pp. 285-301

Smilansky, S. (2000) *Free Will and Illusion*, Oxford U. P.

Smith, A. (2007) 'On Being Responsible and Holding Responsible,' *The Journal of Ethics* 11, pp. 465-84

Smith, M. (1994) *The Moral Problem*, Basil Blackwell.

Sommers, T (2009) 'The Two Faces of Revenge: Moral Responsibility and the Culture of Honor,' *Biology and Philosophy* 24, pp. 35-50

Strawson, P. (1974) 'Freedom and Resentment,' *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

Vargas, M. (2007) 'Revisionism,' *Four Views on Free Will*, ed. J. M. Fischer, R. Kane, D. Pereboom and M. Vargas, Blackwell

Vargas, M. (2008) 'Moral Influence, Moral Responsibility,' *Essays on Free Will and Moral Responsibility*, ed. N. Trakakis and D. Cohen. Cambridge Scholars Publishing (2008), pp. 90-122

Wallace, R. J. (1994) *Responsibility and the Moral Sentiments*, Harvard U. P.

Watson, G. (1996) 'Two Faces of Responsibility,' *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