A Unified Empirical Account of Responsibility Judgments

GUNNAR BJÖRNSSON

Umeå University, University of Gothenburg

KARL PERSSON

University of Gothenburg

1. Introduction[*]

Skeptical worries about moral responsibility go to the heart of our self-conceptions and our understanding of distributive and retributive justice. Unlike few other philosophical issues, they are also widely appreciated and deeply felt by laymen, and this constrains theories about moral responsibility: they should better speak to pre-theoretical concerns, if nothing else to show that these concerns are mistaken.

Appealing to this constraint, both compatibilists and incompatibilists have accused their opponents of constructing notions of responsibility that go beyond or fail to capture what the folk are concerned with.[1] Since claims about folk notions of moral responsibility have an obvious empirical component, philosophers have recently begun taking advantage of the

[*] This paper has benefited from comments from numerous people, including John Eriksson, Caj Strandberg, Felipe de Brigard, Michael McKenna, John Fischer, Dana Nelkin, Josh Knobe, Adam Feltz, the audience at SPP 2009, participants at a research seminar in practical philosophy at the University of Gothenburg and a workshop at UC Riverside, and, not least, an anonymous reviewer for *Philosophy and Phenomenological Research*.

[1] On the incompatibilist side, we have, among others, Cover and O'Leary-Hawthorne 1996: 50-51; Ekstrom 2000: 57; Kane 1996: 16f; 1999: 218, 2002: 310; Pink 2004: 12-14, Smilansky 2000 and Strawson 1986; on the compatibilist side Strawson 1982 [1962]; Dennett 1984: 555-558; and Lycan 2003. Cf. Sommers 2010.

methods of empirical sciences to settle the controversy, focusing mostly on the issue of compatibilism between moral responsibility (or free will) and determinism. The results have been all but straightforward. Not only do efforts to assess whether 'the folk' have compatibilist tendencies suggest that the folk are at least as divided as philosophers, they also suggest that responsibility judgments are affected by a number of factors that philosophers have rarely if ever considered as directly relevant to moral responsibility, such as a target action's tendency to trigger emotional reactions, the level of concrete detail by which the action is described and whether it is set in this world or in an alternative universe, to name a few.[2] A variety of explanations have been suggested for individual results, but we still lack a satisfactory analysis of the everyday concept of moral responsibility that provides a unified explanation of the full spectrum.[3]

In light of the puzzling results, one might suspect that there is no such explanation: perhaps the common sense notion of moral responsibility is fundamentally disjunctive, telling us to apply different criteria for different kinds of scenarios; or perhaps different people have different concepts of moral responsibility.[4] Alternatively, one might suspect that the variety of folk intuitions stems from confusion: perhaps incompatibilist intuitions are due to conflation of determinism and fatalism or conflation of reductionism and eliminativism, or perhaps compatibilist intuitions are due to our general ignorance of the complete causes of our motives and decisions.[5] More charitably, one might suspect that various moral or

[2] Experimental studies of judgments of moral responsibility have recently received attention in *Science* (Nichols 2011), a further indication of the non-esoteric nature of the problem.

[3] See Nahmias 2005, 2007; Nichols and Knobe 2007; Feltz et al. 2009.

[4] See e.g. Knobe and Doris 2010, Feltz et al. 2009.

[5] See Spinoza 1985 [1677]; Nahmias et al. 2007; Feltz et al. 2009.

pragmatic considerations explain why judgments about different cases seem to be guided by different conceptual rules.[6]

To decide whether these suspicions are correct, though, it is not enough to look at intuitions concerned with one narrow range of cases. We need recourse to a comprehensive empirically based theory of folk intuitions about moral responsibility, supported by a wide range of considerations. This is exactly what we hope to provide in this paper. In what follows, we will show how an independently motivated hypothesis about responsibility judgments provides surprisingly straightforward explanations of the more significant results in experimental philosophy. According to this hypothesis—the 'Explanation Hypothesis'—the judgment that an agent is morally responsible for an event is an explanatory judgment saying, very roughly, that a *relevant motivational structure* of the agent is part of a *significant explanation* of the event. Because of how explanatory interests and perspectives affect what we take as significant explanations, this analysis promises to account for the varying tendencies revealed by recent empirical studies, among those our diverging tendencies towards compatibilism and incompatibilism. If it succeeds, there is a unified notion of moral responsibility behind the bewildering variety of intuitions among laymen and philosophers alike.

Elsewhere, we argue that the truth of the Explanation Hypothesis would provide strong reasons to reject incompatibilist or more generally skeptical arguments against moral responsibility. But whether that claim is true or not, the hypothesis provides a richer understanding of folk intuitions and a better basis for distinguishing reliable from mistaken intuitions than the schematic forms of compatibilism and incompatibilism that have been subjected to experimental empirical tests.

[6] See e.g. Nelkin 2007: 256-257 n25; Warmke 2011.

The paper is structured as follows. In section 2, we introduce the Explanation Hypothesis, outlining its basic rationale. In section 3, we indicate some of the phenomena that it accounts for, and some direct empirical support from studies that we have conducted ourselves. In sections 4 and 5, we will look at further empirical data and show how the account of this set provided by the Explanation Hypothesis is either more straightforward or deeper and more detailed than earlier suggestions. In sections 6 we deal with possibly troubling data. In section 7, finally, we take stock and outline how some of the controversies in the philosophical debate about moral responsibility should be understood if the Explanation Hypothesis is correct.

2. The Explanation Hypothesis

The Explanation Hypothesis is best understood as part of a more general account of practices of holding people responsible, of expressing indignation or praise, withholding or increasing cooperative behavior, and distributing punishments and rewards. Generally speaking, humans engage in these practices because we take an interest in providing and confirming social support for our normative expectations and values. This is clearly seen in paradigmatic instances of second-person blame, which involve expressing disapproval to someone who has caused harm or violated normative expectations. It is a striking fact that such disapproval typically is placated by the agent's display of willingness to live up to the expectations, while increased by a steadfast refusal to change. First- and third-personal varieties of blame are equally revealing. In the former, feelings of guilt motivate adjustment of the agent's own behavior and values, and prompt expression of such motivation which, when acknowledged, mitigate feelings of guilt. In the latter, it is clear that the person who is blaming often seeks confirmation of her own disapproval in expressions of similar attitudes from others. Such confirmation is likely to reinforce normative expectations along with behavioral patterns and motivational structures that fit with these expectations.[7] Similar things are true about holding people responsible for good things, by praising or rewarding them for effects, actions, and decisions that help, live up to, or exceed normative expectations.[8]

Consider two facts about these practices of holding responsible. The first is that they need guidance to reliably reinforce the relevant expectations and motivational structures: they need to target the right kind of motivational structure at the right time under the right circumstances. The second is that these practices are in fact directly guided by our judgments of moral responsibility: we tend to hold people responsible only insofar as we think that they *are* responsible. Guiding blame, condemnation, praise, punishments and rewards is arguably the signal psychological function of these judgments. Correspondingly, the signal psychological function of our concept of moral responsibility—of our psychological capacity

[7] For the social and evolutionary function of emotions that dispose humans to engage in pro-social punishment (i.e. of indignation), see e.g. Fehr and Fischbacher 2003, Fehr 2004, and Jaffe and Zaballa 2010. For a nuanced philosophical description of first and second person blame and its psychological role, see Bennett 2002.

The claim here is not that practices of holding people responsible are consciously directed at reinforcing or modifying motivational structures, merely that they are generally structured so as to achieve that goal, be it by instrumental reasoning or instinct shaped by biological or social evolution. It is remarkable that even people whose acts of holding responsible are explicitly aimed at retributive justice rather than the protection of a social system are often highly interested in the public confirmation of support of certain values.

We are also not claiming that our ordinary practices of holding people responsible are *ideal* for reinforcing normative expectations needed for good social life, merely that their performance of this function explains their existence and important aspects of their structure.

[8] The pattern is perhaps less obvious in the positive case, as there is nothing quite like the escalations of blame seen in the second-person case when the agent fails to display guilt or remorse.

to identify cases of moral responsibility—is to guide these practices of holding responsible by guiding our judgments of responsibility.[9]

The suggestion, in light of these facts, is that our concept of moral responsibility plays the role it plays because it tracks conditions under which it *works* to hold someone responsible for something. It is of course well known that responsibility judgments are insensitive to a variety of such conditions. We might judge that agents are responsible for something while thinking that holding them responsible would be counterproductive or pointless, perhaps because of our relations to the agents, or because they are contrarians, or long since dead.[10] But our responsibility judgments do seem to track three general pre-conditions for successfully holding people responsible for something.

First, unless a certain kind of motivational structure has a *general tendency* to be harmful, modifying it is of little use, and unless it has a general tendency to be beneficial, reinforcing that structure or encouraging emulation will do little good. Call this the TENDENCY condition.

Second, unless what we are holding agents responsible for is *straightforwardly and saliently explained* by their motivational structures on the given occasion, in accordance with the general tendency, holding them responsible is unlikely to result in modification or reinforcement of the relevant attitudes. To be ready to modify our ways, we need to see the harm as resulting from a lack of relevant motivational structures; to be strengthened in our

[9] This is not to deny that we can sometimes engage in what might be called ‘blind’ blaming, i.e. primitive negative reactions to violations of social expectations also seen in dogs and many other social animals (tendency to retaliate or create distance, tendency to submissive behavior). Plausibly, our more complex practices and our concept of moral responsibility have developed from such reactions.

[10] See e.g. Smith 2007.

ways, we need to see a valuable outcome as resulting from relevant motivational structures. Call this the EXPLANATION condition.

Our responsibility judgments seem to track both TENDENCY and EXPLANATION. We seem to take people to be morally responsible for bad decisions, actions and other events that are clearly explained by their being overly egoistic, fanatic, ill-willed, careless, inattentive, insensitive, etc., i.e. by motivational structures that generally tend to have bad consequences. We do not seem to target motivational structures that just happen to have bad effects on one occasion, through wayward causal chains. Similarly, we do not take people to be morally responsible for good events unless they are explained by motivational structures in typical ways.

Third, for practices of holding responsible to work, targeted motivational structures must be *of a type that can be reinforced* by holding people responsible for its good effects, *or is amenable to modification* by holding people responsible for its bad effects. Call this the RESPONSE condition.

Again, it seems clear enough that our responsibility judgments tend to track this condition. On the one hand, we normally do attribute responsibility for actions or outcomes that we take to be explained by motivational structures that are relevantly responsive. Most obviously, we hold people responsible for intended good or bad outcomes chosen on the basis of consciously endorsed as well as unconsciously held values. But we also hold people responsible for outcomes resulting from success or failure to recognize or react to morally relevant features of the situation—a person's need, or a risk for harm, say. Both consciously and unconsciously held values and more subtle (in)sensitivities to morally relevant features constitute motivational structures that we take to respond to practices of holding responsible. In particular, we generally assume that holding people responsible for a harm might lead them to better recognize the risk of harm, to consciously strive to avoid harming, and to value safety.[11]

On the other hand, we tend to take compulsion, phobias, severe personality disorders or extreme stress to mitigate responsibility, and we tend to attribute less than full responsibility to children or others who lack the capacity to understand moral values or practices of holding responsible. Under all these latter conditions, actions can be seen as governed by types of motivational structure that are resistant, in various degrees, to practices of holding responsible. (Obviously, types of motivational structure are variously responsive: reasons-based values of normal adults might seem more responsive to practices of holding responsible than, say, tendencies to recognize subtle forms of oppression, or the unarticulated values of small children. This seems to be at least roughly mirrored in the degrees to which we attribute moral responsibility.)[12]

Talk about *moral* responsibility in particular seems to respond to a further condition, namely that the motivational structures in question are *morally significant*, either by having a general tendency to bring about outcomes that are morally significant, or by being based on considerations that are morally significant. Call this the SIGNIFICANCE condition.

[11] That a motivational structure is of a *type* responding to practices of holding responsible does not mean that it would in fact respond in the right way: as already noted, the agent might be dead, or have psychological dispositions that block effects of holding responsible.

[12] One can expect variations in responsibility judgments depending on variations in individuations of types of motivational structures (cf. issues of individuating what Fischer and Ravizza (1998) call 'mechanisms'). But while this makes the implications of RESPONSE somewhat indeterminate, it is not rendered vacuous. Given intuitively plausible assumptions about how people type motivational structures, it seems to make sense of why a variety of conditions pertaining to such structures undermine attributions of responsibility. (For a theory of moral responsibility, rather than an account of *judgments* of moral responsibility like the Explanation Hypothesis, it might be important to specify the *correct* way of typing motivational structures.)

The Explanation Hypothesis says that the folk concept of moral responsibility is sensitive to these four considerations, as follows:

> THE EXPLANATION HYPOTHESIS (EH): People take someone to be morally responsible for an event to the extent that they take it that TENDENCY, EXPLANATION, RESPONSE and SIGNIFICANCE hold.[13]

Furthermore, we might say that EH without SIGNIFICANCE characterizes a notion of 'agential responsibility'. Most people in the moral responsibility debate seem to have such a 'morally neutral' notion of responsibility in mind.[14]

Notice that although EH is meant to capture the considerations that determine people's responsibility judgments, it does not imply that people are aware of these considerations under the descriptions in TENDENCY, EXPLANATION, RESPONSE or SIGNIFICANCE. In particular, it does not imply that people normally conceptualize the relevant motivational

[13] For a more precise formulation of the interaction between the conditions, see Björnsson and Persson *forthcoming* and Björnsson *in progress A*.

[14] See e.g. Fischer and Ravizza 1998: 6; Pereboom 2001: xx. Unlike many accounts of moral responsibility, EH is not primarily restricted to intentions, decisions or actions. The event in question can be of either of those kinds as well as a causal consequence of the agent's action, or something that he failed to prevent. (Given a loose enough understanding of 'event', EH also covers judgments of responsibility for character traits and beliefs.) Though wide in scope, EH falls short of an account of all responsibility judgments. It directly concerns retrospective moral responsibility for events, not judgments concerning prospective responsibility (what it is an agent's responsibility to do or oversee) or concerning whether someone is a responsible agent. (However, since retrospective event responsibility is closely related to these other notions, we expect EH to have implications for them too.) And while it covers holding groups or corporate entities responsible insofar as these can be seen as having motivational structures of the appropriate kind (Björnsson 2011), it might not cover concepts of responsibility governing radically different practices of holding responsible, such as practices of holding

structures in terms of amenability to modification by practices of holding responsible, rather than in terms of capacities to do otherwise, say, or in terms of volitional or rational control. Nor does the broadly consequentialist etiological story that informs EH imply that attributions of responsibility are directly motivated by thoughts about the consequences of holding others responsible. More plausibly, our basic tendencies to hold responsible and responding to being held responsible develop from relatively primitive social emotional tendencies to aggression and social repair. Though these become successively more sophisticated in part because we learn about consequences of directing our emotional reactions in various ways, we continue to be directly governed by thoughts about whether someone brought about or allowed an event in a certain way.[15]

In some ways, EH follows extant compatibilist accounts of moral responsibility (see e.g. Fischer and Ravizza 1998) and extant psychological accounts of judgments of moral responsibility or blame (see e.g. Alicke 2000). TENDENCY and EXPLANATION are meant to explain the almost platitudinal nature of the requirement that agents responsible for an outcome be causally responsible for that outcome.[16] Likewise, RESPONSE and SIGNIFICANCE (in conjunction with TENDENCY and EXPLANATION) are meant to capture the psychological

individuals responsible for what members of their group have done, say, —unless these are reducible to cases of holding groups responsible (cf. Sommers 2009).

[15] Because individual judgments of moral responsibility are typically made without an eye to the efficacy of holding someone responsible, they provide an independent constraint on the practices of holding responsible, sometimes yielding negative verdicts under conditions where the practices would work fine. In Björnsson and Persson *forthcoming* and later in this paper, we explain why skeptical arguments and reflection on determinism are prone to produce such practice-undermining negative judgments. (A neighboring debate concerns whether the *correctness* of responsibility judgments is grounded in the practices or vice versa. For versions of the former view, see e.g. Strawson 1982 and Wallace 1994; for criticism, see Smith 2007.)

[16] Björnsson (2011) argues that EH accounts for one sort of exception to this requirement in relation to collective responsibility.

correlates to the intuitive requirements that responsible agents have a guilty mind, that their behavior is expressive of their moral character, or that they did not act from compulsion. What EH and its etiological underpinnings contribute in the first place is a unified account of these conditions in terms of the psychological role of the concept of moral responsibility. However, our focus here is on the explanations EH offers of a number of disparate and often philosophically puzzling aspects of our judgments of moral responsibility.

In the next section, we will say a little bit more about how the EXPLANATION condition should be understood, and mention some of the phenomena that EH has been used to explain, including results from a recent study designed to directly test the hypothesis. This provides independent motivation for the explanations that we will offer in later sections for recent bewildering empirical results.

3. Applying the Explanation Hypothesis

Many of the most interesting predictions and explanations generated by EH depend on the notion of a *significant* explanation of an event. The key feature of that notion is its selective nature. Suppose that we are asked why a house has just burned down. In answering, we could list a number of conditions, each of which might be a necessary part of a complex sufficient condition for the outcome: there was a thunderstorm; the house was hit by lightning an hour earlier; the house consisted largely of combustible matter; there was oxygen in the air; etc. All of these conditions might be part of a full causal story leading up to the fact that the house burned down, but only a small subset will stand out when we want to give a condensed explanation of that fact. When we do, the fact that the house was hit by lightning will likely grab our attention, whereas the fact that the house consisted of combustible matter or that there was oxygen in the air would be taken for granted as part of what we might call the explanatory 'background'. Typically, the explanatory background consists of conditions that are generally to be expected whereas attention grabbers are conditions that violate such expectations. Generally speaking, we expect houses to be built from some amount of combustible material, and we certainly expect there to be oxygen in the air, but we do not in the same way expect houses to be hit by lightning at some given time.

Everyday explanations are selective in other ways too. The bolt of lightning that hit the house itself had a causal genesis, and there were numerous causal intermediaries between the fact that the house was hit by lightning and the fact that it burned to the ground. These conditions are not likely to be seen as part of the explanans, however. When we explain an event, we cite conditions that provide a particularly telling explanation among those leading up to the event, conditions that satisfy our explanatory interests without immediately raising new and urgent why-questions. If we wonder why the house burned down and are told that the attic insulation caught fire, we will probably wonder *why* the insulation caught fire, and if we are told that there was a separation of positive and negative charges in the neighboring atmosphere, we are likely to ask *how* that explained that the house burned down. By contrast, if we are told that the house was hit by lightning, we will probably be satisfied: we take a house's being hit by lightning to be both the sort of thing that just happens and the sort of thing that causes houses to burn down.

Generally speaking, whether we understand something as a significant or relevant explanation depends on explanatory interests. If we ask a fire engineer why the house burned down, the fact that it was hit by lightening might be part of the explanatory background: perhaps we want to know what was especially important about the construction of the house, or about the fire protection available or missing. Sometimes such interests are naturally explained in contrastive terms: perhaps we want to know why this house burned down *when other houses that were also hit by lightning survived*; in that case, we are looking for some difference between this house and the others that contribute to the full explanation of the event.[17]

When TENDENCY and EXPLANATION in the Explanation Hypothesis refer to a 'significant' explanation, that means an explanation that satisfies our explanatory interests given background assumptions or, differently put, fits our *explanatory frame*. Such a fit is a matter of degree: events stand out from the background to degrees and satisfy our explanatory interests to degrees. The selective and graded nature of significant explanations makes EH a surprisingly powerful theory of judgments of moral responsibility, accounting for a wide variety of otherwise disparate phenomena that were not part of the original motivation for EH.

For example, as we have argued elsewhere, EH accounts for the fact that degrees of external force, threats and costs are seen as variously mitigating moral responsibility.[18] With sufficient external force operating, the agent's motivational structure is irrelevant to explaining his movements: EXPLANATION is violated. Similar for threats. The greater the threat, the more extreme the motivation needed to resist. The agent's motivational structure might still be part of a full explanation of actions even under severe threats, but to the extent that it falls well within what we think that we should expect from people, normatively speaking, we will treat it as part of the explanatory background rather than as significantly

[17] For the selectivity of causal judgments, see e.g. Hart and Honoré 1985, Björnsson 2007, Hitchcock and Knobe 2009.

[18] This seeming platitude is and also confirmed in studies (see e.g. Woolfolk et al. 2006).

explanatory.[19] (Since conformity to normative expectations about motivational structures is a matter of degree, so is the mitigating effect of external threats.)

Similarly, EH accounts for the fact that we sometimes, but not always, take ignorance to mitigate moral responsibility. When we are ignorant of a possible outcome of our actions, the degree to which we care about that outcome will typically not explain these actions. However, sometimes failures to foresee an outcome are fairly straightforwardly explained by agents' motivational structures—perhaps, if the doctor had cared more about risks involved in various procedures, she would have known that the treatment was dangerous. In such cases we might nevertheless attribute responsibility for that outcome.

Interestingly, EH also explains why people can be seen as collectively responsible for outcomes over which they had no individual control. Many think that drivers of gas-guzzling SUVs are morally responsible for effects on the environment, even though no individual SUV's absence would have made a meaningful difference to those effects given the presence of the others. EH provides a straightforward explanation. People who attribute moral responsibility to these drivers think that the effects happened partly because these drivers cared too little about the environment. The explanation they have in mind makes collective reference to the motivational structures of the drivers, thus avoiding the problem that the motivational structure of the individual might have played no significant role taken on its own.[20]

[19] For some effects of normative expectations on causal judgments, see Alicke 1992, Hitchcock and Knobe 2009. Notice that normative expectations do not themselves involve judgments of blameworthiness or praiseworthiness for decisions, actions or outcomes, but concern the motivational states of the agent.

[20] Björnsson (2011) defends this application of EH, arguing that other theories of individual or collective responsibility fail to account for the relevant cases. Notice that the account is not meant to cover the sort of collective responsibility that is sometimes attributed to individuals on the mere ground that they belong to the same group as someone who has caused harm. (That this is a different kind of responsibility is clear from the fact

EH's etiological motivation and apparent capacity to account for central features of everyday thinking about moral responsibility gives it considerable initial plausibility. Further independent support comes from a recent study aimed at testing EH more directly. Subjects were confronted with a number of scenarios and asked to what degree they judged that an agent in the scenario is morally responsible for some event, and to what extent they agreed with an explanation of that event which made reference to the agent's motivation. Answers were indicated on a 0 to 6 scale, where 0 meant 'not at all / strongly disagree' and 6 'fully / strongly agree'. For example:

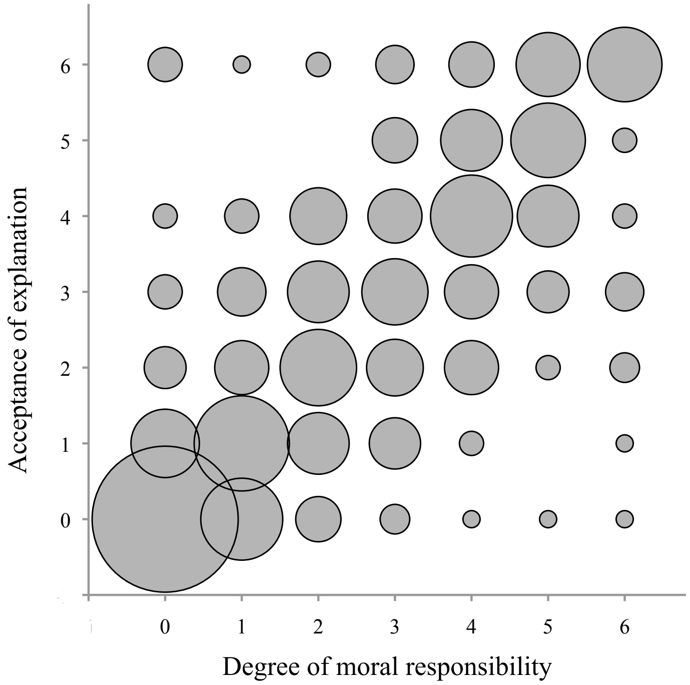
> One hot and sunny day, Kevin went to the store to do some shopping. After parking his car, he noticed a dog in the backseat of the car next to his. Kevin realized that the heat could harm the dog and considered getting help or calling the police, but eventually decided that it was none of his business. While he was shopping, the dog passed out from the heat.
>
> 1. To what degree is Kevin morally responsible for the fact that the dog passed out?
>
> 2. To what extent do you agree with the following explanation of the fact that the dog passed out: *Kevin didn't care enough about the dog*?

that practices of holding responsible that are governed by such attributions of collective responsibility do not target motivational structures of the *individuals* held responsible. Cf. footnote 14.)

Around 30 pairs of answers were collected for each of 15 scenarios. The prediction, based on EH, was that answers in each pair should be significantly correlated, with fairly closely matched values.[21]

This was indeed the result, as visible in Figure 1. The figure represents all 463 pairs of answers, with the area of a bubble centered at a point representing number of pairs at that point, ranging from 1 to 73.[22]

Figure 1



[21] One should not expect a perfect match even on the assumption that subjects understood the questions in the relevant way, since the explanation question might fail to capture the motivational structure that a given subject would see as most explanatory, and since the badness of the outcome might factor into assessments of degrees of responsibility.

[22] In numbers, the correlation (Pearson's r) across all 463 pairs was 0.71 and highly significant (N=463, p <.001); the reliability score (Cronbach's alpha) was 0.83. To test for interaction effects between the questions, other subjects were divided in two groups, each asked only one of the questions. Since the mean values for the two questions were closely matched for the four vignettes in this between-group experiment design, this test did not suggest any significant interaction. For details, see (Johansson 2010).

This result, together with EH's etiological motivation and capacity to explain core features of our thinking about moral responsibility, gives EH the independent support needed to make plausible some of the explanations that we will propose. In the following sections, we will discuss a number of results from studies intended to reveal folk intuitions about moral responsibility, but we begin with philosophically less charged studies concerned with seemingly unrelated aspects of the attribution of blame.[23]

4. Asymmetric responsibilities for known side effects

A number of studies by Joshua Knobe and others suggest that people are significantly more inclined to hold an agent responsible for bringing about bad side effects than for bringing about good side effects when the agent just does not care about these effects.[24] In one study, subjects were presented with either of two scenarios involving a chairman of a board who takes no interest in environmental effects of his decisions. In the first scenario, he knowingly allows a profitable program that *harms* the environment; in the other, he allows a profitable program that *benefits* the environment. Subjects in the harm condition assigned a high degree of blameworthiness (4.8 on a 0 to 6 scale) to the chairman, whereas subjects in the benefit condition assigned a very low (1.4) degree of praiseworthiness (Knobe 2003: 193). This

[23] We do not have space to discuss all empirical studies of responsibility judgments that people have suggested create trouble for standard theories of moral responsibility (see e.g. Knobe and Doris 2010). We have focused on the results that seem most stable, have least obvious explanations, and where different accounts of the concept of moral responsibility make for different predictions. We have excluded discussion of some results that we think EH is particularly well suited to explain, but where other accounts might initially seem to do equally well or better and a satisfactory treatment would require too much space (see e.g. Nahmias and Murray 2010 and Sripada *forthcoming*).

[24] Knobe 2003, Cole Wright and Bengson 2009, e.g.. This and other side-effect asymmetries in attributions of various intentional states or actions are often referred to as 'the Knobe-effect'.

stark difference seems puzzling, as the chairman knew about and was indifferent to the side-effect in both cases.

Since neither Knobe's original study nor follow-up studies in the literature ask explicitly about moral responsibility, one might suspect that the phenomenon has more to do with a contrast between blame and praise.[25] To test this, we ran studies in the same format asking explicitly whether the agent was morally responsible for the outcome. The results were virtually indistinguishable from those in Knobe's study.[26] Moreover, given that moral responsibility is generally seen as a precondition for blameworthiness and praiseworthiness, we might expect an explanation of the responsibility asymmetry to be more basic.

Various explanations of the side effect asymmetry have been suggested, most of which take the effect to rely on our normative understanding of the situation.[27] Appeals to normative expectations will also be part of a full EH account of the phenomenon, but the initial explanation is much simpler. When considering the harm scenario, it is natural to explain the harm to the environment in terms of the motivational structure of the chairman: the environment was damaged because he did not care about such effects. When considering the benefit scenario, however, there is nothing comparable. For example, to say that the environment was helped because the chair did not actively try to harm the environment sounds decidedly less natural. Given EH, this difference in intuitive explanatory judgments straightforwardly accounts for the difference in responsibility judgments.

[25] In general, attributions of blameworthiness, praiseworthiness are correlated but hardly identical. See e.g. Robbennolt 2000, 2588-83, Woolfolk et al 2006, 289-290.

[26] For a scenario very similar to Knobe's chairman of the board scenario, mean moral responsibility for the harm was 5,3 with confidence interval at 95% = 4.9-5.7; for benefit 1.4 with confidence interval at 95% = 0.8-1.9 (33 subjects in each condition).

[27] See Cole Wright and Bengson 2009, Knobe 2010 e.g..

This is enough to show that EH *can* account for the asymmetry in responsibility judgments and, by extension, for the asymmetry in attributions of blameworthiness and praiseworthiness. Moreover, all independent support for EH gives us independent reasons to think that this *explanation asymmetry* is what is actually at the heart of the matter. However, without an independent account of the explanation asymmetry, one might suspect that it depends on the responsibility asymmetry rather than the other way around.

As it happens, there is a rather straightforward account of the explanation asymmetry. In the harm condition, there are two reasons why the chair's lack of environmental concern will seem like a natural significant explanation: (i) people in general take it that indifference towards valuable things often explains harm to them, and (ii) the chair's indifference *stands out* in comparison to our normative expectation that people care appropriately, making it a significant explanation given a typical explanatory frame. Contrast this with the benefit condition. Here, there is no type of motivational structure exemplified by the chair that both (i) has a general tendency to explain outcomes of the relevant sort and (ii) stands out in the required way. For example, there seems to be no general tendency for indifference toward a valuable thing to benefit it, and no general tendency for an interest in profit to lead to benefits to the environment. One could perhaps say that states of not being opposed to X tend to explain why X happens, and also that the chair's not being actively opposed to benefits to the environment was necessary for the beneficial effects. Even granted this, however, the state of not being actively opposed to benefits to the environment conforms strongly to our normative expectations. It thus falls squarely into the explanatory background, along with the presence of oxygen in the company headquarters and the fact that the company had enough resources to run the program.

The suggestion, then, is that the explanation asymmetry and, by extension, the responsibility, blameworthiness and praiseworthiness asymmetries, are due to our normative expectations. Not caring about the environment violates normative expectations and stands out; not being opposed to benefits to the environment falls well within our expectations.[28]

5. Explaining incompatibilist intuitions, and their absence

Most of the philosophical concern with the everyday concept of moral responsibility comes from two features. On the one hand, the concept has a central role in moral thinking and everyday life; on the other, it also seems open to skeptical worries, in particular worries about the compatibility of responsibility with determinism or scientific explanations of our

[28] This explanation is in line with some other explanations that have been proposed. Dana Nelkin (2007, 353) briefly suggests that the responsibility asymmetry might be due to asymmetries in our moral duties. The chair deserves blame in the harm condition because he violates a duty not to harm, whereas the chair in the benefit condition does not deserve praise for the beneficial consequences since it was not his intention to help. That suggestion seems consonant with the explanation developed here, although lacking in details about the connection between duties, blame- and praiseworthiness and moral responsibility. Knobe's own suggestion is that these asymmetries are at bottom a matter of whether people's attitudes fall short of (or exceed) a default set by normative expectations, much as we have argued (Knobe 2010; Pettit and Knobe 2009). But without an account of *how* defaults affect judgments, this seems to be no more than a promising hunch. What we have offered is exactly such an account: defaults affect what stands out from the explanatory background.

One remaining issue is to say how this account generalizes to related asymmetries revealed by other studies, in particular asymmetries in attributions of intentionality and related psychological states. Just as people are willing to attribute responsibility for the effect in the harm condition but not in the benefit condition, people are much more willing to say that the chair *intentionally* harmed, was *in favor of* harming, or *decided* to harm the environment in the former condition than they are to say that he intentionally helped, was in favor of helping, or decided to help the environment in the latter. Similarly, people are somewhat less reluctant to say, in the harm condition, that the chair increased profit *by* harming the environment, or that he harmed the environment *in order to* increase profit, than to say, in the benefit condition, that he increased profit by helping the environment, or that he helped the environment in order to increase profit (Pettit and Knobe 2009, e.g.). It would be surprising if these asymmetries had completely different explanations, but we explain elsewhere how the account provided here extends to these other expressions (Björnsson *in progress B*).

actions. This tendency towards skepticism has been the subject of a flurry of recent experiments, where empirically oriented philosophers have tried to determine whether the folk concept of moral responsibility is incompatibilist or not. The results have been mixed. Much as philosophers, lay people are divided between those who do and those who do not take responsibility to be undermined in deterministic scenarios, or by neurophysiological explanations of decisions.[29] This is interesting in itself, but so is the distribution of incompatibilist intuitions and features that affect these intuitions. We will begin this section by mentioning a number of interesting results that we find particularly clear, and then proceed to explain how they are accounted for by EH.

Most of the studies have been made using an experimental paradigm developed by Shaun Nichols and Joshua Knobe (2007), where subjects are asked questions after having read descriptions of one deterministic and one indeterministic universe. Here is the description of the former:

> Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example one day John decided to have French Fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decision, then it *had to happen* that John would decide to have French Fries. (Nichols and Knobe 2007, 669)

[29] See e.g. Nahmias et al 2005; 2007; Nichols and Knobe 2007; Feltz et al 2009; Roskies and Nichols 2008.

The indeterministic universe, Universe B, is described as just like Universe A, with one exception. Human decision making is said not to be completely caused by what happened before: the decision "did not have to happen"; the agent could have decided differently.

Other, somewhat similar, ways of presenting deterministic scenarios to subjects have been used in various studies, and there is some debate about which is most likely to get across a proper understanding of determinism. Some descriptions might not convey the relevant necessitation of later events by earlier conditions while others, including the description of Universe A above, risk inviting fatalistic rather than deterministic interpretations. For this reason, we see some of the following discussion as provisionary. However, unlike some critics, we do not think that we *now* have grounds for attributing the findings that we will discuss here primarily to misinterpretation of deterministic scenarios.[30]

Here are some of the findings:

CROSS-CULTURAL ROBUSTNESS: When asked whether our universe is more like Universe A than Universe B, between 65% and 85% of subjects in the United States, India, Hong Kong and Columbia have given the indeterminist answer. Moreover, when asked whether, in Universe A, it is possible for a person to be fully morally responsible for their actions, between 63% and 75% have given the incompatibilist answer. Across a variety of different cultural backgrounds, then, a majority of people seem to think that human decisions are not completely caused by the past, and a majority of people seem to think that determinism precludes moral responsibility (Sarkissian et al. 2010). This suggests that incompatibilist intuitions stem from some basic feature of our thinking about moral responsibility. At the

[30] For some possible sources of error, see Sommers 2010, Nahmias and Murray 2010. The worries raised by Nahmias and Murray are based on particularly intriguing empirical results. We suspect that they can be accounted for within the framework provided by EH, but are undertaking studies to test this hypothesis.

same time, a substantial portion of subjects give compatibilist answers, suggesting that although incompatibilist intuitions stem from some basic feature of the relevant concept of moral responsibility, there is no simple and obvious inferential rule that both constitutes competence with the concept and mandates incompatibilist judgments.

COMMITMENT TO MORAL RESPONSIBILITY: Although a substantial proportion of subjects make incompatibilist judgments about deterministic scenarios, few people give both determinist and incompatibilist answers. Moreover, when subjects in a study by Adina Roskies and Shaun Nichols (2008) were asked to assume that our actual universe is deterministic, they were significantly more inclined to ascribe moral responsibility to agents than were subjects considering agents in a merely possible universe. Like the presence of a substantial minority willing to attribute full moral responsibility to agents in Universe A, these results suggests that our everyday understanding of the concept of moral responsibility, although prone to give rise to incompatibilist intuitions, does not *straightforwardly* commit us to incompatibilism. At the very least, it seems that any incompatibilist aspects of the concept can be overridden by a commitment to moral responsibility in the actual world (Roskies and Nichols 2008: 378–387).[31]

CONCRETE VS. ABSTRACT: Studies by Nichols and Knobe (2007) suggest that whether people take agents to be responsible for their actions in a deterministic scenario depends on whether these actions are described abstractly or concretely. Subjects in the 'abstract' condition were asked whether, in Universe A, it is possible for a person to be fully morally responsible for his or her actions, whereas subjects in the 'concrete' condition were asked

[31] The study reported in Nahmias et al 2007: 227 complicates the picture somewhat: the tendency reported by Roskies and Nichols holds for scenarios of psychological determinism, but is reversed for scenarios of neurological or reductionist determinism.

whether a man called Bill is fully morally responsible for intentionally killing his wife and children because he has become attracted to his secretary. In the abstract condition, 86% percent answered 'no'; in the concrete condition, 72% answered 'yes' (Nichols and Knobe 2007: 670, see also Nahmias et al 2007: 227).[32] The puzzle is to explain why people assign responsibility in the concrete but not in the abstract condition when the universe is portrayed the same way in both.

Nichols and Knobe hypothesized that *affect* evoked by the deeds depicted in the concrete case influences judgment, and tested this by presenting 'high affect' and 'low affect' cases presented against a deterministic background scenario. In the high affect case, a man stalks and rapes a stranger; in the low affect case, a man cheats on his taxes. When subjects were presented with the high affect case they where more inclined to ascribe full moral responsibility (64%) than when they were presented with the low affect scenario (23%) (Nichols and Knobe 2007: 675-676). The question, again, is why this is so.[33]

EXPLANATIONS IN TERMS OF MOTIVES VERSUS PRIOR CAUSES: In a study by Adam Feltz and colleagues (Feltz et al. *forthcoming*), subjects read a description of a deterministic world (departing slightly from the Nichols and Knobe paradigm) in which John kills his wife to marry his lover. They were then asked to score their agreement with the claim that John was morally responsible for killing his wife, and later asked to explain John's action. Those who

[32] Judgments of responsibility about concrete actions in a variety of deterministic scenarios seem to yield similar numbers (Nahmias et al 2006).

[33] Results in Feltz 2009 gives some cause for concern about these results. Each subject in their study were confronted with both high affect and low affect scenarios, and 92% gave *the same* answer in both. (According to Adam Feltz (personal communication), the remaining 8%, 4 subjects, attributed responsibility in the high affect case but not in the low affect case.) However, although this study did not display any order effect for the two

explained his actions in terms of John's motives agreed with the attribution of moral responsibility to a significantly higher degree than those who explained his action in terms of events outside John's control. This coincides with how proponents of incompatibilism and compatibilism tend to argue: compatibilists stress the deliberative and motivational features of the agent whereas incompatibilists point to features outside the agent's control. The question is what explains this connection between modes of explanation and compatibilist and incompatibilist intuitions.

PSYCHOLOGICAL VS. MECHANISTIC EXPLANATIONS: Eddy Nahmias, Justin Coates and Trevor Kvaran (2007) provide evidence suggesting that what keeps people from assigning responsibility is reductionism and mechanistic explanations of people's behavior, not determinism (cf. Greene and Cohen 2004). In their experiment, subjects in the 'reductionist' condition were confronted with a deterministic scenario in which *neuroscientists* have discovered the *chemical reactions* and *neural processes in our brains* that completely cause our decisions and actions, and are themselves completely caused by events preceding our births. Subjects in the 'non-reductionist' condition were instead confronted with a scenario in which *psychologists* had discovered the *thoughts, desires and plans in our minds* that completely cause our decisions and actions, and are themselves caused by events preceding our births.

When subjects were asked to what extent they agreed that people should be held responsible for their actions if the relevant form of determinism were true, the level of agreement was significantly lower among subjects in the reductionist condition: 41% agreed at least somewhat, as compared to 89% in the non-reductionist condition (Nahmias et. al.

scenarios, it also did not prevent subjects from reading both scenarios before answering and adjusting their answers to achieve consistency (ibid. 8–9).

2007: 227).[34] The question is why many, though not all, take reductionist determinism to reduce or be incompatible with moral responsibility.

EH explains all these results in terms of how different explanatory frames determine what strikes us as straightforward explanations. As spelled out in section 3, what we take to straightforwardly explain an event depends on our explanatory interests and background assumptions. Under normal circumstances, we think of decisions, actions and outcomes of actions as straightforwardly explained by agents' motivational structures, as we employ a central folk-psychological explanatory model. However, deterministic scenarios make salient abstract explanatory frames in which agents' motivational states seem comparatively insignificant: all events are explained or 'fully caused' by 'what happened before', and no particular event in the unfolding universe is more significant than any other. Or more precisely: the only ordering of explanatory significance is that *earlier* events explain later events, making the initial state of the universe the most significant explanans. In everyday contexts, where our explanatory interests are more concrete, more proximal explanations are often much more significant, as prior causes are less straightforwardly connected to the explanandum. Recall: the fact that lightning struck the house provides a more straightforward explanation of why the house burned down than does the fact that there was a separation of positive and negative charge in the neighboring atmosphere, even if the latter explains the lightning.

The effect of introducing the deterministic scenario is thus to strongly invite use of the more abstract explanatory frame in which agents' motivational structures are not seen as significant explanations of their actions. Insofar as we accept that invitation while judging

[34] The difference was considerably lower when asked about a foreign planet inhabited by creatures whose lives and societies resemble ours, and also considerably lower when asked about concrete actions.

whether agents in the deterministic scenario are responsible for their actions, EH predicts that we will no longer see these agents as responsible. This explains why subjects who are asked whether agents can be fully morally responsible in Universe A tend to give a negative answer.[35] However, although the deterministic scenario invites an abstract perspective, subjects might resist the invitation, instead employing folk-psychological explanatory models that account for decisions and actions in terms of desires, values and preferences. Since folk-psychological models are likely to be easily accessible, it is not surprising that the incompatibilist tendency, although strong, is limited.

For these reasons, we think that sensitivity to explanatory frames can explain why judgments of moral responsibility are subject to strong tendencies towards incompatibilist intuitions, without incompatibilism being part of the concept of moral responsibility. But it also provides explanations of the other results listed above.

First, to the extent that similar practices of holding responsible are part of all the cultures in which intuitions have been tested using the Nichols and Knobe paradigm, and given the etiological story motivating EH, we should expect subjects in all these cultures to employ a concept of the same sort.[36]

Second, given the central everyday role of explanatory models in which motivational structures are significant and the central social role of practices of holding responsible, we should expect a psychological commitment to attributions of moral responsibility. We should

[35] It also explains the strong tendency to give positive answers when asked the same question about agents in Universe B: in one study, between 89% and 95% subjects said that agents in Universe B could be fully morally responsible (Nichols and Knobe 2007: 676). Since decision-making in this universe is said not to be completely caused by prior events, it cannot be fully explained within the abstract frame. This invites reliance on our most salient, folk-psychological, explanatory models for explanations of actions and decisions.

[36] However, cultures with quite different practices of holding responsible might have quite different concepts of responsibility. See Sommers 2009.

thus not expect people who think that the world is deterministic to make their responsibility judgments employing the abstract deterministic model: that would involve distancing oneself from one's dispositions to blame and praise people in the actual world.[37]

Third, the EH explanation of incompatibilist intuitions also explains why questions about agents' responsibility for *concrete* actions are less likely to trigger incompatibilist intuitions. Given a concrete action and perhaps even a concrete motive, folk-psychological explanatory models that are dependent on such details will provide more salient alternative to the abstract deterministic explanatory model.

Fourth, EH provides a straightforward explanation of why subjects more readily ascribe responsibility in 'high affect' cases than in 'low affect' cases. Since a serious transgression strongly violates our expectations, it is exactly the sort of event that calls for an explanation, and more precisely an explanation telling us *why it happened here but not in cases where the expectations are satisfied*. The explanatory model provided by the deterministic scenario cannot answer that call as it provides the same abstract explanation for cases that violate and cases that satisfy expectations: both kinds are fully caused by prior events. By contrast, an appeal to the agent's motivational structure provides a salient possible explanation: the violation happened in this case because the agent didn't care appropriately about the relevant values, in the way people typically do when they satisfy expectations. What we suggest, then, is that 'high affect' cases trigger responsibility judgments even among subjects primed with an abstract deterministic explanatory model because they call for explanations that can only be given by abandoning the abstract model. Since the weaker transgressions of 'low

[37] Of course, such distancing is part of certain religious and spiritual practices, but these are avowedly esoteric.

affect' cases stand out less, however, their call for a corresponding explanation in terms of motivational structures will be less strong.[38]

Fifth, the EH explanation of incompatibilist intuitions straightforwardly predicts that people who explain actions in terms of agents' reasons, thoughts and desires will attribute higher degrees of responsibility than those who explain actions in terms of other sorts of causes.

Finally, EH explains why a reductive, neurological, deterministic scenario undermines judgments of responsibility much more strongly than a non-reductive, psychological deterministic scenario. Again, the reason is that the former is more likely to prompt explanatory frames that do not invoke the motivational structures of the agent. In the reductive scenario, we are not only lead to adopt a new set of explanatory categories (chemical and neurological processes) but also, perhaps, to discard our everyday explanations of human action (depending on whether reductionist explanations are seen as conflicting with those), whereas the explanatory categories postulated in the non-reductive scenario coincide with those employed in folk-psychological explanations.[39]

[38] Explanations of the 'affect' asymmetry proposed by Nichols and Knobe (2007: 671-73) give affect a central explanatory role. An explanation based on EH could do that too, if affect prompted focus on the agent's motivational structures over and above that effected by violations of normative expectations. However, there is some empirical evidence suggesting that affect plays no major role in explaining compatibilist intuitions about 'high affect' cases (see Cova, Bertoux et al. *forthcoming*), evidence favoring non-affective accounts like the one proposed here.

[39] This explanation also tells us why the distinction between a psychological and a neurological explanation has no effect on attributions of moral responsibility when, unlike in the Nahmias et al. (2007) studies, the two types of explanation are similarly suggestive as to the agent's RESPONSE-satisfying motivational structures (De Brigard et al. 2009).

What we have seen in this section is how EH can explain a variety of results coming from the empirical study of folk skepticism or incompatibilism about moral responsibility. These explanations complement EH-based explanations that we have given elsewhere of why various philosophical arguments seem to undermine moral responsibility, and of why certain replies to these arguments seem effective (Björnsson and Persson *forthcoming*).

6. Objection: Identification and responsibility

EH predicts that attributions of responsibility are constrained by matters of control. For an agent to be seen as responsible for her actions, they must be seen as explained in normal ways by relevant motivational structures. However, Joshua Knobe and John Doris have suggested that recent studies by Woolfolk, Doris and Darley (2006) show that moral responsibility does not require agential control:

> Woolfolk, Doris and Darley … ran a series of experiments in which subjects were given short vignettes about agents who operated under high levels of constraint. In one such vignette, a character named Bill is captured by terrorists and given a 'compliance drug' to induce him to murder his friend:
>
> > Its effects are similar to the impact of expertly administered hypnosis; it results in total compliance. To test the effects of the drug, the leader of the kidnappers shouted at Bill to slap himself. To his amazement, Bill observed his own right hand administering an open-handed blow to his own left cheek, although he had no sense of having willed his hand to move. The leader then handed Bill a pistol with one bullet in it. Bill was ordered to shoot Frank in the head...

> The researchers then manipulated the degree to which the agent was portrayed as identifying with the behavior he has been ordered to perform. Subjects in one condition were told that Bill did not want to kill Frank; those in the other condition were told that Bill was happy to have the chance to kill Frank. The results showed that subjects were more inclined to hold Bill morally responsible when he identified with the behavior than when he did not. In other words, people assigned more responsibility when there were higher levels of identification even though the agent's behavior was entirely constrained. The study therefore provides strong evidence for the view that people are willing to hold an agent morally responsible for a behavior even when that agent could not possibly have done otherwise (Knobe and Doris 2010).

In this case, it not only seems clear that Bill could not have done otherwise; it also seems perfectly clear that his motivational structures played no normal role in explaining his actions.[40] If Knobe and Doris are right that these studies undermine the idea that responsibility requires the ability to do otherwise, then, they should equally undermine EH's EXPLANATION requirement. Instead, what seems to be required is merely that the agent behaves *in accordance with* how she wants to behave.[41]

[40] In the terminology of Fischer and Ravizza (1998: 31), he lacked not only guidance control, but also regulative control.

[41] This would fit with the familiar idea that an agent is acting 'of her own free will' in the sense relevant for moral responsibility if the motivational structures that govern her action are ones that the she wants to be governed by: this is exactly what is missing in the unwilling addict or compulsive kleptomaniac. (Much contemporary work in this tradition departs from Frankfurt 1971.) The twist, of course, is that in this case, the governing motivational structures are those of someone else.

However, we think that a closer look at the results from the Woolfolk study shows that Knobe and Doris have seriously exaggerated their importance. The mean responsibility score in the 'high identification' condition where Bill was happy with the way things unfolded was 3.25 on a 1 to 7 scale. That might seem like a substantial degree of responsibility, but there are strong reasons to doubt that this says anything that undermines either the requirement of alternative possibilities or the EXPLANATION requirement.

First, when subjects were asked to assess whether Bill was 'free to do other than he did' on a 1 to 7 scale, the mean was 1.98 (Woolfolk et al. 2006: 296–297). This suggests that subjects did not uniformly see the compliance drug as an *absolute* constraint, making it possible for them to think that Bill's motivational structures played *some* role in explaining why Bill shot Frank. Moreover, the mean score in the high identification condition was only 1 unit lower than in the 'low identification' condition where Bill did *not* want to kill Frank: 3.25 rather than 2.25 on the 1 to 7 scale. If subjects did not see the compliance drug as an absolute constraint, this comparatively small difference could well have resulted from the attribution of *some* explanatory significance to Bill's motivational state in the high identification condition.

Second, although subjects were asked about Bill's responsibility *for Frank's death*, they might have ascribed some degree of responsibility in order to express disapproval of Bill's endorsement of the events in the high identification condition. This is not a mere theoretical possibility: in our own studies, subjects' free form explanations of their responsibility attributions show that quite a few confound responsibility for outcomes with responsibility for decisions preceding the outcomes, leading them to attribute responsibility for outcomes entirely beyond the agent's control unless given the opportunity to target the decision specifically.[42]

For these reasons, we do not think that the Woolfolk results lend weight to the conclusion that Knobe and Doris wants to draw, or to threaten EXPLANATION.

7. Reframing the debate

We should stress again that the discussions in the preceding sections is somewhat tentative. Some of the empirical results that we have discussed come from studies with relatively few subjects, and some of the questionnaires used might contain unfortunate formulations. At this time, however, we have not seen convincing positive reason to doubt that the phenomena that we have tried to explain are real. Moreover, each of the explanations provided here gains considerable support from EH's capacity to account for wide variety of other aspects of our thinking about moral responsibility. This is not to deny that more studies are needed to work out in greater detail how responsibility judgments interact with explanatory frames, but we think that such studies can be fruitfully guided by EH. For example, we think that it will be fruitful to look more closely at various further factors that affect explanatory judgments and see whether they affect responsibility attributions. Similarly, it will be interesting to see to what extent different intuitive ways of individuating motivational structure types might affect attributions of moral responsibility, as EH predicts. Given EH, it is also a highly interesting question to what extent laymen are variously committed to specific explanatory frames in ways that make their intuitive judgments of moral responsibility resistant to the various framing effects that we have considered.

[42] See Johansson 2010: 14-18.

Before conducting such further studies, however, we have enough reason to think that EH is roughly correct to ask about its implications. Obviously, EH does not in itself tell us what intuitions are correct, incorrect, based on confusion, or reliable. What it does provide, however, is a more substantial basis for discussing such issues. Assuming, as we think reasonable, that eliminativist worries about basic folk psychological explanations are incorrect, there clearly are explanatory frames given which agents' RESPONSE-satisfying motivational structures are part of significant explanations of decisions, actions and outcomes. Given EH, this means that the only way to produce incompatibilist intuitions, except through fallacious reasoning, is by shifting the explanatory frames of people attributing moral responsibility away from these basic folk psychological frames. In section 5, we argued that this is what deterministic scenarios tend to do. Elsewhere, we argue that regress arguments, manipulation arguments, and arguments from luck work in the same way.[43]

All this suggests that the debate between compatibilists and incompatibilists should itself be reframed as a debate about what the relevant explanatory frames are for judgments of moral responsibility that govern our practices of holding people responsible. Moreover, since traditional arguments for or against incompatibilism rely on intuitions of responsibility that depend on these very explanatory frames, we have reason to think that entirely different kinds of arguments would have to be adduced to settle the issue.

[43] Björnsson and Persson *forthcoming*, Björnsson *in progress C*.

Bibliography

Alicke, Mark D (1992) Culpable Causation. *Journal of Personality and Social Psychology* 63, pp. 368–78.

Alicke, Mark D (2000) Culpable Control and the Psychology of Blame. *Psychological Bulletin*, 126, pp. 556–74.

Bennett, Christopher (2002) The Varieties of Retributive Experience. *The Philosophical Quarterly*, 52, pp. 145–63.

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

Björnsson, Gunnar (2011) Joint Responsibility without Individual Control: Applying the Explanation Hypothesis, in *Compatibilist Responsibility: Beyond Free Will and Determinism*, eds. Jeroen van den Hoven , Ibo van de Poel and Nicole Vincent, Springer 2011, pp. 181–199.

Björnsson, Gunnar (in progress A) Illusions of Undermined Responsibility.

Björnsson, Gunnar (in progress B) The Explanation Explanation of Side-Effect Effects.

Björnsson, Gunnar (in progress C) The Manipulation in Manipulation Arguments.

Björnsson, Gunnar and Persson, Karl (forthcoming) The Explanatory Component of Moral Responsibility. *Noûs*, early view, DOI: 10.1111/j.1468-0068.2010.00813.x

Cole Wright, Jennifer and Bengson, John (2009) Asymmetries in Judgments of Responsibility and Intentional Action. *Mind & Language*, 24, pp. 24–50.

Cova, Florian; Bertoux, Maxime; Bourgeois-Gironde, Sacha and Dubois, Bruno (forthcoming) Judgments About Moral Responsibility and Determinism in Patients with Behavioural Variant of Frontotemporal Dementia: Still Compatibilists! *Consciousness and Cognition*.

Cover, J.A. and O'Leary-Hawthorne, John (1996) Free Agency and Materialism. In J. Jordan and D. Howard-Snyder (eds), *Faith, Freedom, and Rationality*. Lanham, MD: Roman and Littlefield.

De Brigard, Felipe, Mandelbaum, Eric and Ripley, David (2009) Responsibility and the Brain Sciences. *Ethical Theory And Moral Practice*, 12, 511–24.

Dennett, Daniel (1984) I Could Not Have Done Otherwise: So what? *Journal of Philosophy*, 81, pp. 553–565.

Ekstrom, Laura W (2000) *Free Will: A Philosophical Study*. Westview.

Fehr, Ernst and Fischbacher, Urs (2003) The Nature of Human Altruism. *Nature*, 422, pp. 137–40.

Fehr, Ernst (2004) Don't Lose Your Reputation. *Nature*, 432, pp. 449–50.

Feltz, Adam; Cokely, Edward T. and Nadelhoffer, Thomas (2009) Natural Compatibilism versus Natural Incompatibilism: Back to the Drawing Board. *Mind & Language*, 24, pp. 1–23.

Feltz, Adam; Perez, Ashley and Harris, Magan (forthcoming) Free Will, Causes, and Decisions: Individual Differences in Written Reports. *Journal of Consciousness Studies*.

Fischer, John and Ravizza, Mark (1998) *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge U. P.

Frankfurt, Harry G (1971) Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, 68, pp. 5–20.

Greene, Joshua.; Cohen, Jonathan (2004) For the Law, Neuroscience Changes Nothing and Everything. *Philosophical Transactions: Biological Sciences*, 359, pp.1775–85

Hart, H. L. A. and Honoré, Tony (1985) *Causation in The Law.* Oxford U. P.

Hitchcock, Christopher and Knobe, Joshua (2009) Cause and Norm. *Journal of Philosophy* 106:11, 587–612.

Jaffe, Klaus and Zaballa, Luis (2010) Co-Operative Punishment Cements Social Cohesion *Journal of Artificial Societies and Social Simulation* 13 (3) 4, <http://jasss.soc.surrey.ac.uk/13/3/4.html>

Johansson, Erik (2010) Testing the Explanation Hypothesis using Experimental Methods. Bachelors thesis, Linköping University. URI: urn:nbn:se:liu:diva-57308

Kane, Robert (1999) Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. *Journal of Philosophy*, 96, 217–240.

Knobe, Joshua (2003) Intentional Action and Side Effects in Ordinary Language. *Analysis* 63, pp.190–93.

Knobe, Joshua (2010) Person as Scientist, Person as Moralist. *Behavioral and Brain Sciences* (2010), 33, pp 315-329

Knobe, Joshua and Doris, John (2010) Responsibility. In *The Moral Psychology Handbook*, ed. John Doris. Oxford: Oxford University Press, pp. 321-54.

Lycan, William (2003) Free Will and the Burden of Proof. In A. O'Hear (Ed.), *Minds and Persons: Royal Institute of Philosophy Supplement* (pp. 107–122). Cambridge, England: Cambridge University Press.

Nahmias, Eddy; Morris, Stephen; Nadelhoffer, Thomas and Turner, Jason (2005) Surveying Freedom: Folk Intuitions about Free Will and Moral Responsibility. *Philosophical Psychology*, 18, pp. 561–84

Nahmias, Eddy; Morris, Stephen; Nadelhoffer, Thomas and Turner, Jason (2006) Is Incompatibilism Intuitive? *Philosophy and Phenomenological Research*, 73, pp. 28–53.

Nahmias, Eddy; Coates, Justin and Kvaran, Trevor (2007) Free will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions. *Midwest Studies in Philosophy*, 31, pp. 214–42

Nahmias, Eddy and Murray, Dylan (2010) Experimental Philosophy On Free Will: An Error Theory For Incompatibilist Intuitions. In Jesús Aguilar, Andrei Buckareff and Keith Frankish (eds.) *New Waves in Philosophy of Action*, Palgrave Macmillan, pp. 189-216.

Nelkin, Dana K (2007) Do We Have a Coherent Set of Intuitions about Moral Responsibility? *Midwest Studies in Philosophy*, 31, pp. 243–59.

Nichols, Shaun (2004) The Folk Psychology of Free Will: Fits and Starts. *Mind & Language*, 19, pp. 473–503

Nichols, Shaun (2011) Experimental Philosophy and the Problem of Free Will. *Science*, 331, pp. 1401–03.

Nichols, Shaun and Knobe, Joshua (2007) Moral Responsibility and Determinism: the Cognitive Science of Folk Intuitions, *Noûs* 41:4, 663–685

Pereboom, Derk (2001) *Living without Free Will*. Cambridge University Press.

Pettit, Dean and Knobe, Joshua (2009). The Pervasive Impact of Moral Judgment. *Mind & Language* 24:5, 586–604.

Pink, Thomas (2004) *Free Will: A Very Short Introduction*. Oxford University Press.

Robbennolt, Jennifer K (2000) Outcome Severity and Judgments of 'Responsibility': A Meta-Analytic Review1. *Journal of Applied Social Psychology*, 30, pp. 2575–609.

Roskies, Adina L. and Nichols, Shaun (2008) Bringing Responsibility Down to Earth. *Journal of Philosophy*, 105, pp. 371–88.

Sarkissian, Hagop; Chatterjee, Amita; Brigard, Felipe De; Knobe, Joshua; Nichols, Shaun and Sirker, Smita (2010) Is Belief in Free Will a Cultural Universal?. *Mind & Language*, 25, pp. 346–58.

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

Smilansky, Saul (2000) *Free Will and Illusion*. New York: Oxford U. P.

Sommers, Tamler (2009) The Two Faces of Revenge: Moral Responsibility and the Culture of Honor. *Biology and Philosophy* 24, pp. 35–50.

Sommers, Tamler (2010) Experimental Philosophy and Free Will. *Philosophy Compass*, 5, pp. 199–212.

Spinoza, Baruch (1985) [1677] *The Collected Works of Spinoza,* Princeton U. P.

Sripada, Chandra Sekhar (forthcoming). What Makes a Manipulated Agent Unfree? *Philosophy and Phenomenological Research*.

Strawson, Peter F (1982) [1962] *Freedom and Resentment* in *Free Will* ed. Gary Watson. Oxford U. P., pp. 59–80.

Strawson, Galen (1986) *Freedom and Belief*. Oxford: Clarendon

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

Warmke, Brandon (2011) Moral Responsibility Invariantism. *Philosophia,* 39, pp. 179-200.

Woolfolk, Robert L.; Doris, John. M. and Darley, John M (2006) Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility. *Cognition* 100, 281–301.