Joint Responsibility without Individual Control: Applying the Explanation Hypothesis

Gunnar Björnsson
Linköping University
University of Gothenburg

Abstract:
This paper introduces a new family of cases where agents are jointly morally responsible for outcomes over which they have no individual control, a family that resists standard ways of understanding outcome responsibility. First, the agents in these cases do not individually facilitate the outcomes and would not seem individually responsible for them if the other agents were replaced by non-agential causes. This undermines attempts to understand joint responsibility as overlapping individual responsibility; the responsibility in question is *essentially* joint. Second, the agents involved in these cases are not aware of each other's existence and do not form a social group. This undermines attempts to understand joint responsibility in terms of actual or possible joint action or joint intentions, or in terms of other social ties. Instead, it is argued that intuitions about joint responsibility are best understood given the *Explanation Hypothesis*, according to which a group of agents are seen as jointly responsible for outcomes that are suitably explained by their motivational structures: something bad happened because they didn't care enough; something good happened because their dedication was extraordinary. One important consequence of the proposed account is that responsibility for outcomes of collective action is a deeply normative matter.

1 Joint moral responsibility without individual control

Sometimes a number of individuals seem *jointly* morally responsible for events over which they, as individuals, had no control. Consider a simplified case:

> *The Lake*: Alice, Bill and Cecil each have a small boat in East Lake outside their town. One day last spring, each painted the boat and, unknown to the others, poured excess solvent into the lake. In the back of their heads, they all knew that this could affect the wildlife, but each of them decided that it would be a hassle to dispose of the solvent in a safe way and hoped that nothing bad would happen. However, as the

> solvent from all three diffused throughout the lake over the next few days, its concentration became high enough everywhere to prevent micro-organisms in the lake from reproducing during the next few weeks, thus leaving higher organisms without food and effectively wiping out all fish in the lake. The concentration of solvent exceeded the threshold for the microorganisms by quite some margin: although the solvent from only one of the three would not have been enough to kill off the fish, the solvent from two would have.

Let us assume that all three agents satisfied conditions of moral accountability. They were not being forced or manipulated to do what they did and they had both the capacity to reason and reflect on the values involved and the relevant sort of control over their own decisions and actions. Then it seems that we can rightly hold them responsible for recklessly pouring solvent into the lake. But to just about everyone that I have confronted with the case, it also seems clear that they are morally responsible *for the death of the fish*, that is, for an outcome of their actions over which they had no control as individuals. Similarly, it seems that voters can be morally responsible for the outcome of a referendum, citizens for toppling a dictatorial regime, consumers for good or bad practices of companies they patronize, and frequent flyers and drivers of SUVs for climate effects, even though, as individuals, they could not have significantly affected those outcomes, practices or effects.

The question of this paper concerns the conditions for such joint responsibility for outcomes of collective actions. In the next section, I explain why a case like *The Lake* provides difficulties for standard ways of understanding collective responsibility. In section three, I propose a preliminary analysis of joint responsibility based on variations on *The Lake*. To support this analysis, section four introduces the *Explanation Hypothesis*, a model of our concept of moral responsibility that was developed to account for various aspects of individual moral responsibility for decisions, actions and outcomes. In section five, I show how the Explanation Hypothesis subsumes and deepens the analysis of section three. In section six, finally, I suggest a way of turning the Explanation Hypothesis' characterization of our *concept* of moral responsibility into an account of moral responsibility as such. One of the important consequences of the proposed account is that responsibility for outcomes of collective action is a deeply normative matter.

Some caveats are in order. First, the concern of this paper is *moral, retrospective responsibility for events*. Space prevents me from saying anything about the tight and interesting connections between this topic and other questions discussed under the heading of

"responsibility"—questions concerning legal liability, moral or legal obligations to *ensure* outcomes or to *take* responsibility for outcomes by compensating those harmed, and questions about what characterizes responsible persons, or responsible decision procedures. Second, since the concern is with joint responsibility of *individual* agents, I will not say anything about the claim that collectives can be responsible for an outcome when *none* of their members are. (For recent defences of "autonomous" corporate responsibility, see Arnold 2006, Pettit 2007, Tännsjö 2007, Copp 2007; for criticism see Corlett 2001, Haji 2006, McKenna 2006, Miller 2007.) Third, the primary concern here is with *outcome* responsibility rather than responsibility for decisions. The conditions under which individuals are responsible for their decisions are themselves highly contestable, but I will assume that all individuals in the cases discussed are autonomous, in control of their own decisions and actions, capable of rational deliberation, suffering from no motivational maladies, and, as a result, responsible for their own acts or failures to act. Fourth, since our concern is with difficulties pertaining specifically to the understanding of how individuals are *jointly* responsible for outcomes, I will assume that other difficulties pertaining to outcome responsibility can be overcome, in particular the fact that outcomes often depend on factors outside the agent's control. (For discussion, see Feinberg 1968: 681-82; Nagel 1976; Sverdlik 1987: 74; May 1992: 42-45; Enoch and Marmor 2007 e.g.). Finally, although it is clear that individuals can be jointly responsible for good outcomes, I will follow most of the literature and focus on responsibility for *bad* outcomes. It should be clear, however, that the discussion generalizes to good outcomes.

2 Difficulties

As we shall see, neither the standard notion of individual responsibility for outcomes, nor typical strategies for making sense of collective moral responsibility explain the intuition that the agents in cases like *The Lake* are responsible for the outcomes in question.

On a standard conception, an individual agent is morally responsible for a harm to the extent that some morally faulty aspect of her behaviour played a significant causal role in producing that harm (Feinberg 1968: 674; May 1992: 15). The difficulty is to see how the reckless acts of the agents in *The Lake* play a significant causal role.

We have already noted that neither agent made any difference to the survival of the fish given the other acts, so significance cannot require such *difference making*. On the other hand, the solvent contributed by each agent was causally involved in bringing about the outcome.

But causal *involvement* cannot in itself be what accounts for individual responsibility for the collective outcome. Suppose that there are two solvents. Solvent X works as before, preventing microorganisms from reproducing, but it can do so by means of either of two distinct but equally powerful chemical processes, X1 and X2, depending on whether solvent Y is present. Solvent Y is itself incapable of doing any damage except in extreme concentrations, but will favour process X2 in the presence of solvent X. Suppose further that whereas Bill and Cecil poured solvent X into the lake, Alice contributed solvent Y, thus slightly changing the way the solvents from Bill and Cecil prevented micro-organisms from reproducing. Then it is not clear that she would be morally responsible for the outcome.

Intuitively, it might seem that the relevant causal involvement would have to be one of at least *facilitating* the causal process, or make it more likely to produce the outcome (cf. Petersson 2004). But while that might be true for responsibility for outcomes of individual actions, it is not required in *The Lake*. Suppose that when the concentration of solvent reaches above what would be provided by two polluters, the process by which the microorganisms are prevented from reproducing is both slowed down and made more open to possible disturbances, thus slightly decreasing the objective probability of the outcome. Then it is true of each of the polluters that he or she actually (but unwittingly) lowered the probability that the fish would die and obstructed that process to some degree, given the actual contribution from the other two. Nevertheless, the three polluters would still seem to be jointly responsible for the death of the fish; it still died because of their actions.

Now consider the corresponding case with only one agent involved:

> *Adam's Lake*: Because of rare but naturally occurring processes, a poisonous substance is produced in the mud at the bottom of the lake. The amount would be just enough, by itself, to prevent the microorganisms from reproducing. Over the same period, Adam is painting his boat, recklessly pouring excess solvent into the lake that contains the very same poisonous substance. The overall result is that the lake contains more than enough to kill off the microorganisms. In fact, at this concentration, the processes preventing the reproduction are a little slower than they would have been if Adam had not disposed of his solvent this way. In the end, though, the microorganisms are wiped out.

Though it is clear that Adam is morally responsible for recklessly pouring solvent into the lake, most people seem reluctant to say that he is responsible for the death of the fish. At the

very least, it was much clearer that Alice, Bill and Cecil were so responsible in *The Lake*. This strongly suggests that the responsibility attributed to the three is fundamentally collective. *Taken together*, the faulty behaviours of Alice, Bill and Cecil clearly played a significant causal role in wiping out the fish; *individually*, they did not.

The problem posed by *The Lake* for standard accounts of responsibility for outcomes of individual action is equally a problem for attempts, like that of Stephen Sverdlik (1987), to reduce collective or shared outcome responsibility to individual outcome responsibility. But it also poses a problem for standard attempts to understand forms of collective or shared responsibility, whether reductive or not. Since the most obvious cases in which we hold agents responsible for an outcome *as a group* are cases where they have either worked together towards some goal or failed to do so, such attempts are often cast in terms of actual or possible joint agency or joint intentions (Held 1970, Rescher 1998, Kutz 2000, Miller 2006, Sadler 2006, Shockley 2007, e.g.). Less obvious and more controversial are cases where members of a community are responsible for outcomes of acts by other members because members empower and are empowered by each other, and thus "shares in what each member does, and … should feel responsible for what the other members do" (May 1992: 11).

The Lake fits neither of these patterns. Since Alice, Bill and Cecil performed their acts independently and without knowledge of the others, they had no intentions to act together with the others. Nor is it likely that our ascription of joint responsibility relies on the assumption that they could reasonably have formed such intentions. Moreover, we have no reason to think that they form a group the members of which empower each other. For all we know, they might see each other as enemies. Still, they seem jointly morally responsible for the death of the fish.

What is clear from *The* Lake and similar examples is that a number of individuals can be jointly responsible for an outcome if, *together*, they play a significant causal role for that outcome. Structurally, this relation between the actions of the individuals and the outcome is similar to well-known attempts to analyse causes, not as necessary conditions or difference makers, but as non-redundant parts of nomically sufficient conditions for effects (Mackie 1974; cf. Wright 1988). In *The Lake*, the actions of the three agents are pair-wise sufficient for the outcome, each action being a non-redundant part of such a pair. It might thus be tempting to explain the joint responsibility of the three agents in such terms (Braham and van Hees *ms*). Unfortunately, any such attempt will run into deep problems with cases of what

David Lewis (1986b) calls "causal preemption". Suppose that instead of pouring solvent into *East Lake*, Alice built a contraption that monitored the concentration of solvent in the lake and set it to empty her bucket of solvent into the lake should the level not rise high enough to kill the fish. Since Bill and Cecil contributed enough solvent, Alice's contraption was never triggered. In this case, she clearly would not be responsible for the outcome, even though her action would be a non-redundant part of sufficient conditions for the death of the fish (a condition that included her action and the contribution of either Bill or Cecil).[1]

Elsewhere I have defended a way for theories of causation dealing in sufficient conditions to adequately account for cases of causal preemption (Björnsson 2007). But something more would need to be said even with such an account at hand. The fact that Adam poured solvent into the lake was a non-redundant part of a sufficient condition for the death of the fish, together with the fact that some volume of mud at the bottom of the lake emitted the same amount of poisonous substance; yet Adam's responsibility for that outcome is much less obvious than Alice's, Bill's and Cecil's in *The Lake*. Apparently it matters whether the actions of other agents are involved; the fundamental problem of joint responsibility is *why*. This is where I hope to make progress.

3 A preliminary analysis of responsibility for outcomes of collective action

To understand joint responsibility, the first thing to be clear about is the required relation between the collective and the outcome for which they are responsible. As a first approximation, what is required seems to be that, together, the responsible agents play a significant role in the *explanation* of the outcome: the fish died *because of* Alice, Bill and Cecil. With some qualifications, this is very much in line with the idea that individual outcome responsibility requires that the individual's behaviour played a significant causal role in the outcome. However, talk about *causal* (as opposed to *explanatory*) role suggests that the responsible parties *brought about* or *produced* the outcome rather than merely *letting it happen*, and we know that production is not required for outcome responsibility:

[1] Other problems are provided by probabilistic case where there are no causally sufficient conditions for outcomes, and so-called "switching" cases, where necessary parts of sufficient conditions seem to change the way an outcome happens without being causally responsible for it (cf. the case where Alice contributes solvent Y). These are also problems for counterfactual analyses in the tradition of David Lewis (1973); for discussion, see e.g. (Collins et al 2004; Björnsson 2007).

> *The Well:* Eric, Fiona and George are spending a Sunday afternoon in the woods, each thinking that he or she is the only person within miles. Suddenly they hear cries for help coming from an area with especially dense vegetation. Although the cries are disturbing and continues for a long while, each ignores them while thinking that they could be part of a prank, or that whatever might be going on is none of their business. Had they walked in the direction of the cries, however, they would have found a woman, Hannah, who had accidentally fallen into a partially overgrown old well but was hanging onto a ledge a meter or so down, screaming for help and slowly losing her grip. Since no one came to her help, Hannah eventually fell down into the dried up well and died as she hit the rocks at the bottom. The story could have ended differently, however. One person would not have been able to pull her up without help, but had any two of those who heard her cries come to her rescue, they would have been able to save her.

It seems that if they learned the truth of what happened, Eric, Fiona and George could rightly blame themselves for not having investigated the call closer. But it also seems that they are to some extent morally responsible for the fatal outcome of the accident (though not, of course, for the accident itself), and they certainly seem responsible for the fact that Hannah wasn't saved. They could have saved her, but they did not. As in *The Lake*, the responsibility involved seems to be essentially collective. In a version of *The Well—Esther's Well*—Esther is the only person in place to hear Hannah's cries. Like Eric, she ignores the cries for dubious reasons; like Eric she would have been unable to save Hannah even if she had responded. But whereas Eric, Fiona and George seemed clearly responsible for the fact that Hannah wasn't saved, Esther clearly is not. *Esther's Well* highlights the essentially joint nature of Eric's, Fiona's and George's responsibility in *The Well*, just as *Adam's Lake* did in relation to *The Lake*.

In *The Well*, unlike in *The Lake*, there is a sense in which none of the three were *involved* in the process leading to the final outcome: indeed, it seems that they could all have been absent and nothing in that process would have been different (ignoring minute differences in the gravitational field and the like). Nevertheless, it seems that their inaction *explains why* Hannah wasn't saved. This is the notion of "explaining why" that seems relevant for our ordinary attribution of moral responsibility in these cases.

Thus far I have suggested that *the agents* should play a significant role in the explanation of the outcome. But more needs to be said about the required sort of involvement. As we have

already seen from *The Well*, the relevant involvement need not consist of any particular sort of positive *intentional action*: perhaps Eric was sitting on a rock, Fiona climbing a tree, and George running across a meadow instead of helping Hannah. Similarly, no *decisions* on part of members of the group need to be involved in the explanation. Perhaps none of the three even considered the possibility of finding out whether they could help; perhaps they just noted, absent-mindedly, that someone seemed to be in need of help but failed to see any reason to take action. That would not seem to remove their responsibility as long as they could have considered the possibility to help, and would have done so if they had cared more about the needs of others. That no decision is needed can be made even clearer with a case involving negligent ignorance where there is no awareness of risk involved. Suppose that Alice, Bill and Cecil poured the solvent into the lake while being unaware of its lethal potential. They could still be responsible for the outcome if the reason they were unaware was that they lacked concern for the environment or for taking in relevant information, and if that explained why they failed to react to the warning signs on the cans of solvent.

In all these variations, we might say that some morally "faulty" aspect of behaviour explains the outcome, but the behaviour seems faulty only because it is explained by the wrong sensitivity to values, or the wrong motivational structure. If Alice, Bill and Cecil were ignorant of the solvent's lethal potential due to other factors than a lack of appropriate concern, their responsibility for the death of the fish is undermined. Similarly, suppose that George was wearing headphones and did not hear Hannah's cries for help. Or suppose that he heard the cries and started walking towards the well but was trapped by impenetrable vegetation blocking his way and delaying him until it was too late. In neither case would he seem to be responsible for the outcome. The best explanation for that, it seems, is that in these cases, unlike in the original scenario, George's concern or lack of concern fails to explain why he didn't reach the well in time.

Another thing to notice is that the outcome needs to be explained by the motivational structure in a "normal" way. If Dave finds out that Alice, Bill and Cecil lack appropriate concern for the environment and draconically proceeds to poison their lake to teach them a lesson, their lack of concern might be part of the explanation of the death of the fish in the lake, but they are not thereby morally responsible for it. Similarly, if George's lack of concern for others had made him ignore a discussion of feasible paths through the forest, and if as a result he was stuck in the mud and unable to heed Hannah's call, it is not clear that he is thereby morally responsible for not having come to her rescue.

Judging from the variations of *The Lake* and *The Well*, it seems that the two groups of people are responsible for the outcomes because the outcomes are explained (in a "normal" way) by the agents' motivational structures. The fish died because Alice, Bill and Cecil lacked appropriate concern for the environment; Hannah's accident had a fatal outcome because Eric, Fiona and George lacked appropriate concern for their fellow human beings. The same seems to hold for cases of moral responsibility for *good* outcomes. Suppose that each member of a trio discovers and mends a leaking sewer out of concern for the environment and that the reduction of pollution secured by any two of them would have been enough to save the fish in the nearby lake, but not the reduction secured by only one agent. Then it would seem reasonable to say that the fish survived because these three individuals cared about the environment, and they would seem to be correspondingly (jointly) responsible for that outcome.

The question remains, however, whether we can expect this analysis to survive still further variations, and whether it generalizes to other cases of collective responsibility. Moreover, we have yet to explain why the *individuals* are jointly responsible for the outcomes, given this diagnosis. It is one thing to say that the group is responsible, another to say that the members of the group are, and it might be thought that attributions of moral responsibility in cases like these involve some kind of mistake. Perhaps our desire to hold someone responsible prompts us to *confusedly* assign joint responsibility for outcomes on the ground that (a) each individual is responsible for wrongfully risking some bad effect—an adverse environmental effects, say—and (b) what they risked actually took place because of these wrongdoings, taken collectively. The suspicion that there is something amiss with our judgments gains force from a comparison of Alice's responsibility in *The Lake* and Adam's in *Adam's Lake*. In spite of performing identical actions the upshots of which are causally involved in bringing about the death of the fish in the same way, and in spite of the fact that their actions result from identical motivational structures, Alice's responsibility was *much* clearer than Adam's. And in spite of acting in the very same way as Esther for the very same reasons and having exactly the same possibility to save Hannah—none—, only Eric seemed responsible for the fact that Hannah wasn't saved. This is bound to strike some readers as arbitrary.[2]

[2] See (Zimmerman 1985, p. 116-17) for an argument that seems to assume that differences of this sort cannot make for different degrees of responsibility.

I address these issues in the next three sections. Section four introduces an independently motivated hypothesis about our concept of individual retrospective moral responsibility, the *Explanation Hypothesis*. In section five, I explain how it subsumes the analysis of joint responsibility developed in this section. This gives us reason to think that our present analysis will generalize to further cases. Moreover, it suggests that the different attributions of responsibility to Alice and Adam are no more arbitrary than attributions of outcome responsibility in general. Although the Explanation Hypothesis is primarily an empirical hypothesis about our concept of responsibility, supported by its predictive power, it strongly suggests an account of moral responsibility. In section six, finally, I introduce that account—*Explanatory Responsibility*—and discuss how it makes issues of outcome responsibility deeply normative.

4 The Explanation Hypothesis

In two recent papers (Björnsson and Persson 2009, *ms*), Karl Persson and I have argued that a wide variety of intuitions about individual responsibility for decisions, actions and outcomes can be explained if we understand our concept of moral responsibility as shaped by our interest in holding people responsible. What follows is a brief and simplified version of that story.

People hold each other responsible for a variety of events in a variety of ways. We blame or express indignation towards people who have brought about or failed to prevent something bad for lack of proper concern, and praise or express moral admiration towards those who have brought about or let happen something good at remarkable costs to themselves. Sometimes our expressions of so-called "reactive" attitudes are as simple as a frown or a smile. At other times we are more elaborate, punishing or demanding explanation or compensation, or distributing rewards and honours. And we direct analogues of all these reactions towards ourselves.

Our interest in holding people responsible is largely an interest in shaping motivational structures—values, preferences, behavioural and emotional habits, etc—in order to promote or prevent certain kinds of actions or events that we like or dislike. Consciously or unconsciously, we often hold ourselves and each other responsible for various outcomes so that we will behave responsibly and take into account possible outcomes of the sort that we have been held responsible for. This is not to deny that we often hold people responsible for reasons of desert, without an eye to deterring or encouraging agents or third parties. The claim

is merely that general reformative interests very much drive and shape our practices of holding people responsible. (For instance, consider the way expressions of indignation are placated when agents express regret and real motivation to avoid repeats, and consider plausible evolutionary rationales for our reactive attitudes.)

In order for our practices of holding people responsible to reliably affect outcomes in this way, they need to be targeted at motivational structures of types that are a) systematically tied to those outcomes and b) tend to be amenable to modification when targeted by these practices, and need to be so when instances of the motivational structure type c) explain the outcome in a salient straightforward way that supports learning.

Undoubtedly, our concept of moral responsibility plays a central role in determining whom to hold responsible for what. In particular, expressions of indignation and requests for explanation are withheld when we conclude that the putative target of these practices was not responsible for the objectionable decision, action or outcome. Since our concept of moral responsibility plays this role, it would not be surprising if it has been shaped by the need to identify proper targets for our practices of holding people responsible, identified by conditions a) through c) above.[3]

This provides motivation for what we call the "Explanation Hypothesis", an empirical hypothesis about the conditions under which we take people to be responsible for some event:

> *The Explanation Hypothesis:* People take P to be morally responsible for E to the extent that they take[4] 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)*: Type M motivational structures are part of a reasonably common sort of significant explanation of type O outcomes.

---

[3] In connecting moral responsibility to reactive attitudes and practices of holding responsible, this hypothesis is closely related to a category of accounts starting with Peter Strawson's (1962) paper "Freedom and Resentment". In (Björnsson and Persson *ms*) we indcate how our particular way of spelling out this connection avoids some of the standard objections raised against such accounts.

[4] In saying that people "take" GET, RR and ER to hold, I do not mean that they are consciously aware of the considerations defined by these conditions in making their judgments of responsibility under these descriptions, only that judgments are in fact determined by such considerations.

Reactive Response-ability (RR): Type M motivational structures tend to respond in the right way to agents being held responsible for realizing or not preventing type O outcomes.

Explanatory Responsibility (ER): S is part of a significant explanation of E of the sort mentioned in GET.

My focus here will be on the two explanatory requirements, GET and, in particular, ER, but a few words are needed to avoid misunderstanding of RR. It is meant to capture the idea that certain types of motivational structures are impervious to blame, praise or other practices of holding people responsible, and that this undermines moral responsibility. RR thus explains why we typically take moral responsibility to be diminished when behaviour is driven by compulsion, phobias, severe personality disorders and extreme stress.

Since RR concerns how motivational structures respond to blame, praise, etc., it is easy to think that the Explanation Hypothesis understands judgments of moral responsibility as forward-looking, concerned with whether holding someone responsible would reform her behaviour. That would be a misunderstanding, however. The fact that someone's motivational structure is of a *type* that tends to respond in the right way does not mean that it is likely to do so in this case. A particular instance of a type that tends to respond appropriately might be resist reform: disdain might satisfy RR, but disdain for morality might be self-protecting. Moreover, various extraneous factors might mask the motivational structure's disposition to react in the right way: perhaps the agent is disposed to react adversely to criticism, say, or perhaps she suffered from a stroke immediately after her action and no longer has the cognitive capacity to understand what she is held responsible for. To be directly forward-looking, judgments of moral responsibility would have to be sensitive to such masks, but they clearly are not; they are essentially backward-looking, concerned with what *explained* the outcome in question.

Among motivational states and outcomes that satisfy RR, there are basically two kinds of explanation that also satisfy GET: First, events are often explained by the fact that we want them sufficiently, as our desires guide our goal-directed cognitive mechanisms ("The trial was all due to Dr. Ortega's relentless passion for justice"; "Her tragic death was due to Mr. Inza's obsession with revenge"). Second, the fact that we do not sufficiently want something not to happen often explains why we let it happen ("The new factory was allowed to pollute the river because the CEO didn't care about the environment"; "He missed his daughter's game

because he cared more about his work than about her").[5] Consequently, we take people to be responsible for a bad outcome when we think that it happened because they wanted them ("Mr. Inza is to blame for her death") or because they didn't care enough to prevent them ("The pollution is the CEO's fault"), and take people to be responsible for a good outcome when it happened because they wanted it ("Dr. Ortega deserves all credit for the trial").[6]

According to the Explanation Hypothesis, our everyday concept of an *explanation why something happened* is at the core of our thinking about moral responsibility. One key feature of that concept is that it is highly *selective*. Suppose that a house has just burned down and that we are asked why. In answering, we could list a number of conditions, each of which might be a necessary part of 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.[7] All of these conditions, and countless more, 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.

---

[5] It is an interesting question whether GET-satisfying explanations require awareness on part of the agent that the sort of outcome in question might take place or whether it can be enough that the person would have been aware and acted on the information if the person had possessed a different motivational structure. We are currently investigating this, and preparatory studies suggest that most people come down on the latter side. For some of the philosophical controversy, see (Zimmerman 2008, ch. 4; Sher 2009).

[6] It is possible that GET should be restricted to these two broad kinds of explanation.

[7] In (Björnsson 2007) I argue that our causal reasoning is *primarily* directed towards sufficient rather than necessary conditions and that this is explained by the connection between causal thinking and instrumental reasoning: instrumental reasoning is primarily directed at ensuring certain states of affairs rather than making them possible. The priority of sufficiency over necessity explains why causation is compatible with many varieties of overdetermination and ultimately explains why responsibility is not a matter of difference making. (All this simplifies matters by ignoring probabilistic causation and explanation.)

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.

Our everyday notion of explanation is selective in another way 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 provide explanations of an event, we cite a condition that we take to provide a particularly *telling* explanation among those leading up to that event, a condition that satisfies 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 neighbouring 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.

When condition ER in the Explanation Hypothesis refers to a *significant* explanation, that means an explanation that satisfies our explanatory interests and background assumptions or, differently put, fits our *explanatory frame*. The selective nature of significant explanations makes the Explanation Hypothesis a surprisingly powerful account of judgments of moral responsibility. Obviously, the hypothesis can account for the fact that we take people to be responsible for most intended outcomes of their actions: because of our powerful goal-directed mechanisms, such outcomes are straightforwardly explained with reference to what we want to achieve, and most of our everyday preferences satisfy RR. But relying on the selective nature of significant explanations it also provides a unifying account of how a wide variety of otherwise disparate phenomena affect judgments of responsibility. As I have argued elsewhere (Björnsson and Persson 2009; *ms*), it explains why we take it that (a) external force, (b) threats and (c) ignorance mitigate moral responsibility to various degrees, as well as why we take it that (d) those who actively participate in the production of an outcome have a higher degree of responsibility for it than those who merely allow others do it, that (e) someone who takes initiative is more responsible than someone who tags along, and that (f) agents are more responsible for known negative than for known positive side-effects that the agent does not care about. It also explains why judgments of responsibility tend to be

undermined by considerations suggesting that (g) our decisions are a matter of luck, (h) our actions are, ultimately, the upshots of events over which we have no control, or (i) our behaviour can be given reductionistic, mechanistic explanations, as well as why (j) the felt conflict between determinism and moral responsibility is lessened when people consider concrete cases, and especially cases involving grave moral transgressions.

5 The Explanation Hypothesis and collective responsibility

The explanatory power of the Explanation Hypothesis, along with its etiological motivation, gives us reason to think that the everyday concept of retrospective moral responsibility has a structure that straightforwardly incorporates our preliminary analysis of joint responsibility in section three: In cases of joint responsibility, the motivational structures of all participants are seen as parts of a significant explanation of the outcome. This gives us independent reason to expect further cases of joint responsibility to conform to the same analysis, thus providing a first answer to the generalization worry.

More specifically, the Explanation Hypothesis explains both why we take the agents of *The Lake* to be responsible for the death of the fish and why we take them to be *jointly* responsible. We see them as *responsible* for the outcome because the three conditions GET, RR and ER are satisfied for each of them, and we see them as are *jointly* responsible because their motivational structures are part of a significant explanans only taken together with the motivational structures of the other two.

Start with the last claim. Compare the following two answers to the question: why did the fish in the lake die?

(1) Alice, Bill and Cecil didn't care about the environmental effects of their actions.

(2) Alice didn't care about the environmental effects of her actions.

Whereas (1) sounds like a perfectly good explanation, (1) is clearly problematic, for two reasons. First it brings attention to the fact that Alice's carelessness made no difference to the outcome because there would have been enough solvent in the lake without it, and although difference making doesn't always undermine explanatory claims it might do so in this case.[8]

[8] The model of causal judgment developed in (Björnsson 2007) explains the restricted role of difference making or counterfactual dependence in causal judgments and shows why the lack of

But (2) is also problematic because it focuses on Alice at the exclusion of Bill and Cecil who played exactly the same role in killing off the fish. Both these defects are absent in (1). That the trio didn't care about the environmental effects of their actions straightforwardly explained why they poured solvent into the lake, and the resulting concentration of solvent explained why the fish died. Of course, not all their actions or all the solvent was needed for that outcome, but there is no privileged subset of these actions that would provide a better explanans. For example, if we explained the death of the fish by mentioning the carelessness of Alice and Bill, we would misleadingly suggest that Cecil had less to do with the outcome than the other two. For that reason, such a restricted explanans would not provide us with an acceptable straightforward explanation.

Now consider the claim that the motivational structure of *each* agent satisfies GET, RR and ER for the outcome in question. First, it satisfies GET because the outcome is explained by a lack of concern to avoid that sort of outcome in the normal way. The most common explanation of this type will be one in which an *individual's* lack of concern explains the outcome, but we frequently explain outcomes in terms of attitudes of members of a group: "The kids next door play loud music because they don't care about the neighbours"; "Sweden rejected the Euro because many Swedes were afraid of losing political independence"; etc. Second, the motivational structures also satisfy RR: we have assumed that the individuals involved satisfy conditions needed for individual responsibility for decisions and action. Finally, we have just seen that the individual agent's motivational structure satisfies ER, as it is alluded to in the joint explanation given by (1).

Contrast this case with *Adam's Lake*. Just like Alice's lack of environmental concern, taken on its own, Adam's lack of concern does not itself strike us as straightforwardly explaining the death of the fish. But whereas Alice's is *part* of a significant explanation that satisfies ER, expressed in (1), it is not clear that Adam's is. For example, the following answer to the question of why the fish died in *Adam's Lake* seems strained:

(3) Adam didn't care about the environment and a poisonous substance was produced at the bottom of the lake.

Although both conjuncts mention conditions that are part of a complete causal explanation of

---

counterfactual dependence might undermine the claim that Alice's carelessness caused or explained the death of the fish in the lake. This effect would be even stronger in the version of *The Lake* where her contribution actually lowered the probability of the outcome.

the death of the fish, their conjunction does not form the most *salient* explanation of the outcome. It would be considerably more natural to appeal to the fact that the lake was poisoned, as the causes of the poisoning are diverse. Moreover, among those causes, the fact that a poisonous substance was produced at the bottom of the lake would be seen as more significant than Adam's contribution, since it actually made a difference to the outcome.

Intuitions about *The Well* are explained almost exactly as intuitions about *The Lake*. Eric, Fiona and George are seen as jointly responsible for the fact that Hannah wasn't saved because it is naturally explained with reference to *their* lack of concern, but not with reference to, say, Eric's lack of concern in particular. The defect of an explanation singling out one individual is more strongly marked than in *The Lake*. "Why wasn't Hannah saved?" "Because Eric didn't care to see whether he could help!" The answer invites the reply that Eric couldn't have saved Hannah on his own, and does so even more strongly than (1) invited the reply that the fish would have died without Alice's action: at least Alice's action was directly causally involved in blocking the reproduction of the microorganisms whereas Eric's inaction made no definite difference at all.[9] (This explanatory inadequacy is of course even more accentuated in *Esther's Well*, where Esther's lack of care clearly does not explain why Hannah wasn't saved.)

What we have seen, then, is how the Explanation Hypothesis supports the diagnosis of joint responsibility provided in section three. Given that so many other aspects of our thinking about moral responsibility is well understood given this account, we should expect further variations on the cases discussed here to conform to the same pattern.

For similar reasons, we should hesitate before saying that typical intuitions about cases like *The Lake* result from confusedly attributing joint responsibility based on (i) *individual responsibility for decisions and actions* and (ii) *non-distributive collective responsibility for an outcome*, that is, collective responsibility that does not imply corresponding responsibility

[9] The Explanation Hypothesis also implies that subtle differences in characterizations of outcomes might yield different verdicts about moral responsibility. It is intuitively clear that Eric, Fiona and George are responsible for the fact that Hannah wasn't saved, but it is less clear that they are responsible for her death. If we ask why she wasn't saved, it is natural to cite, say, the trio's lack of concern, but if we ask why she died, it is considerably more natural to cite the fact that she fell into an old well or didn't watch where she was going than to cite our non-intervention. Different explananda yield different explanatory frames: unlike the fact that she died, the fact that she wasn't saved implies that she was in danger, thus relegating her initial fall into the well into the explanatory background.

for members of the collective. The argument given here suggests that intuitions of joint responsibility rely on the same sort of considerations as do intuitions about individual responsibility. From the point of view of our concept of retrospective moral outcome responsibility, then, the attribution of joint responsibility is in no way arbitrary. Nor is it arbitrary, from an etiological point of view, that we should have a concept that yields this pattern of judgments; a focus on cases with a straightforward explanatory connection between suitable motivational structures and outcomes is crucial for the sort of moral reform that much of our everyday practice of holding people responsible is aimed at. One might worry, though, that it is *unfair* that Alice should be held responsible (together with Bill and Cecil) for the death of the fish whereas Adam is not, given that both were equally reckless and contributed solvents that were similarly causally involved in processes leading to the death of the fish. But this is a familiar problem for outcome responsibility in general, not specifically for joint responsibility or for the analysis proposed here. Factors outside the control of an agent are part of what determines the outcome of her behaviour: only one of two equally reckless drivers is responsible for the death of a child, because only one had a child run out into the street in front of him; only one of two equally courageous and skilled lifeguards is responsible for having saved a life, because only one had the opportunity.

One prediction of the Explanation Hypothesis is that people might be seen as jointly rather than individually responsible for an outcome even in cases where each individual could have prevented the outcome. Think of a version of *The Well* where any one of Eric, Fiona and George could have saved Hannah using a winch next to the well. We might still be reluctant to say that *Eric* is responsible for the fact that Hannah wasn't saved because it arbitrarily picks out Eric at the exclusion of the other two. The significant explanans is still that *none of the three* cared enough to go see whether help was needed; that corresponds to the most natural assignment of responsibility, namely jointly, to all of them.

Another prediction, borne out by almost every discussion of distributive collective responsibility, is that we will ascribe joint responsibility in many cases where agents act together, with joint intentions, since these tend to be cases where agents' motivational structures are involved in straightforwardly explaining the intended outcome. Similarly, intuitions about corporate responsibility bear out the prediction that we will ascribe moral responsibility for outcomes to corporations (organizations, nations, clubs) insofar as we take them to have structures that both straightforwardly explain their actions or omissions and corresponding outcomes and are open to modification by practices of holding these

corporations responsible (see e.g. French 1984; May and Stacey 1991).

For both cases of joint action and cases of corporate moral responsibility, the Explanation Hypothesis predicts attributions of quite different degrees of responsibility to different members of a collective that are causally involved in producing or failing to prevent some outcome. For example, we might think that a stream has been polluted because a certain company doesn't care about the environment, but we do not thereby think that the janitor at the company headquarters is responsible for the pollution. He might have somehow facilitated the process leading to the pollution, but his motivation is not thereby part of a *significant* explanation in the way that the motivational structures of the CEO or members of the board are likely to be. And the same might be true about a member of the board who voted against the polluting activity, or even about someone who voted for it because she thought that that was the way to minimize the harm by allowing her to minimize the resulting pollution.

6 Explanatory Responsibility and the normativity of retrospective outcome responsibility

As we have seen, the Explanation Hypothesis promises a unified account of our judgments of individual and collective responsibility, an account that sees our ascription of essentially *joint* responsibility in cases like *The Lake* or *The Well* as integral to our thinking about moral responsibility in general. Moreover, although it does not say what the relation of moral responsibility *is*, it strongly suggests such an account. Given the Explanation Hypothesis' account of our concept of moral responsibility, it might seem reasonable to simply assume that the relation of moral responsibility corresponds to what is identified when the concept is applied without any mistakes, that is, when GET, RR and ER hold.

Things are not quite that simple, however, because the selective nature of our explanatory judgments makes them sensitive to differences in explanatory frames. For example, it seems that when people are encouraged to abstract away from the level of detail that we employ in everyday explanations of actions and to focus on causal factors outside our control, they are less inclined to find motivational structures explanatorily significant, and less inclined to ascribe responsibility (Björnsson and Persson 2009, *ms*). In the same way, explanatory judgments often depend on *normative* expectations or ideals. Suppose that that a child falls and breaks an arm during some rough and tumble play. A person who thinks that mothers ought to be strongly protective of their children is more likely to explain this fact with reference to the mother's lack of protective concern, and thus more likely to take the mother

to be responsible for the accident.[10]

If the Explanation Hypothesis captures out concept of moral responsibility and if there is a determinate, objective, truth of the matter as to whether people are moral responsible for certain outcomes, the "significant explanations" referred to in GET and ER needs to be restricted. The most obvious way to do so is to require that they are significant relative to a *correct explanatory frame*: relative to *correct* normative ideals, *correct* background assumptions, and *relevant* explanatory interests and explanatory perspectives. "Objectifying" the Explanation Hypothesis, we would thus get the following characterization of moral responsibility:

> *Explanatory Responsibility:* P is morally responsible for E to the extent that E is an outcome of a type O and P has a motivational structure S of type M such that GET, RR and ER hold relative to a correct explanatory frame.

Obviously, Explanatory Responsibility only implies determinate judgments of responsibility given substantial assumptions about what the correct explanatory frames are. This is not the place to defend some such assumptions,[11] but the fact that moral responsibility would depend on the correctness of normative expectations is itself a highly significant consequence.[12] Because of it, fundamental issues in normative ethics are directly relevant to questions of moral responsibility.

As an example, consider how issues of joint responsibility are affected by the disagreement about the existence of reasons to do one's own part in a cooperative scheme even when others are known not to, or to "keep one's own hands clean". Thus far, I have discussed cases where, for all the agents knew, their acts could have made a difference individually to the outcome for which they are responsible. Moreover, this feature might seem essential to the cases. For example, if Alice had poured solvent into the lake knowing for sure that it would make no significant environmental difference or even slowed down ongoing

---

[10] For empirical data illustrating some effects of normative expectations on explanatory judgments, see e.g. (Alicke 1992), (Knobe and Fraser 2008), (Hitchcock and Knobe *ms*) and (Sytsma et al *ms*).

[11] In (Björnsson and Persson *ms*) we argue that explanatory frames of the sort that motivate most of our everyday judgments of moral responsibility should be preferred to the frames that are induced by sceptical arguments against moral responsibility.

[12] For related discussions of how normative aspects affect judgments of responsibility, see (Smiley 1992).

damage, that could clearly undermine her responsibility for the death of the fish as her contribution would no longer be explained with reference to a lack of care. But suppose that there are moral reasons for people to do their part in appropriate cooperative schemes that do not depend on the possibility of actually significantly furthering the ultimate point of these schemes. Then people might be jointly responsible for bad outcomes that they, as individuals, *knew* they could not prevent: if they had all been more concerned to do their part, the outcome would have been different.

If there are non-consequentialist reasons of this sort, their strength will also have major impact on what we are responsible for. Given high enough normative expectations that people should avoid working for or purchase the goods of organizations that are responsible for certain bad outcomes, it will seem that great many people without direct causal influence on these outcomes are nevertheless responsible for them, i. e. for such things as the effects of a company's environmental policy, the persecution of members of organized labour in undemocratic countries, or the enactment of severe oppression of civilians on occupied territories. After all, if people had cared more and been more "principled", many such things could have been very different. This in turn raises difficult questions about the relation between normative expectation and psychological realism: since it seems unlikely that people will live up to these expectations under present circumstances, are they really reasonable? If correct, the Explanation Account makes clear just how such questions are central to issues of collective responsibility, by being directly relevant for the identification of significant explanations.[13]

---

[13] Earlier versions of this text have been presented and received valuable input at the *International Conference on Moral Responsibility* in Delft, August 2009, the Centre for Applied Ethics at Linköping University, the Department of Political Science and the Department of Philosophy, Linguistic and Theory of Science at University of Gothenburg, and the Department of Philosophy, Lund University. I am also grateful to participants at the CEU 2009 summer school on moral responsibility, and to comments from Ibo van de Poel and an anonymous reviewer for this volume.

Bibliography

Alicke, M. D. 1992. Culpable Causation. *Journal of Personality and Social Psychology* 63: 368-378

Arnold, Denis G 2006. Corporate Moral Agency. *Midwest Studies In Philosophy* 30: 279-291.

Björnsson, Gunnar 2007. How Effects Depend on Their Causes, Why Causal Transitivity Fails, and Why We Care about Causation. *Philosophical Studies* 133: 349-390.

Björnsson, Gunnar and Persson, Karl 2009. Judgments of Moral Responsibility: A Unified Account. *Society for Philosophy and Psychology*, 35th Annual Meeting 2009 PhilSci archive. http://philsci-archive.pitt.edu/archive/00004633/

Björnsson, Gunnar and Persson, Karl *ms*. The Explanatory Component of Moral Responsibility. Forthcoming in *Noûs*

Braham, Matthew and van Hees, Martin *ms*. An Anatomy of Moral Responsibility. *Manuscript*

Collins, John; Hall, Ned; and Paul, L. A. (eds) 2004: *Causation and Counterfactuals*. The MIT Press.

Copp, David 2007. The Collective Moral Autonomy Thesis. *Journal of Social Philosophy* 38: 369-388.

Corlett, J. Angelo 2001. Collective Moral Responsibility. *Journal of Social Philosophy* 32: 573-584.

Enoch, David and Marmor, Andrei 2007. The Case Against Moral Luck. *Law and Philosophy* 26: 405-436.

Feinberg, Joel 1968. Collective Responsibility. *The Journal of Philosophy* 65: 674-688.

French, Peter A. 1984. *Collective and Corporate Responsibility*. Columbia U. P.

Haji, Ish 2006. On the Ultimate Responsibility of Collectives. *Midwest Studies in Philosophy* 30: 292-308.

Held, Virginia 1970. Can a Random Collection of Individuals be Morally Responsible?. *The Journal of Philosophy* 67: 471-481.

Hitchcock, Chrisopher and Knobe, Joshua *ms*. Cause and Norm. Forthcoming in *Journal of Philosophy*

Knobe, J., Fraser, B. 2008. Causal Judgment and Moral Judgment: Two Experiments. *Moral Psychology* Vol 2, ed. Sinnott-Armstrong, W., 441-447. MIT Press.

Kutz, Christopher 2000. *Complicity*, Cambridge U. P.

Lewis, David 1973. Causation. *Journal of Philosophy* 70: 556–567. Reprinted in Lewis 1986a, 159–172.

Lewis, David 1986a. *Philosophical Papers*, Vol. II. Oxford U. P.

Lewis, David 1986b. Postscripts to "Causation". In Lewis 1986a, 172–213.

Mackie, John 1974. *The Cement of the Universe*. Clarendon Press.

May, Larry 1990. Collective Inaction and Shared Responsibility. *Noûs* 24: 269-277.

May, Larry 1992. *Sharing Responsibility*. University of Chicago Press.

May, Larry and Hoffman, Stacey (eds.) 1991. *Collective Responsibility: Five Decades of Debate in*

Theoretical and Applied Ethics. Rowman & Littlefield.

McKenna, Michael 2006. Collective Responsibility and an Agent Meaning Theory. *Midwest Studies in Philosophy* 30: 16-34.

Miller, Seumas 2006. Collective Moral Responsibility: An Individualist Account. *Midwest Studies In Philosophy*, 30: 176-193.

Miller, Seumas 2007. Against the Collective Moral Autonomy Thesis. *Journal of Social Philosophy* 38: 389-409.

Nagel, Thomas 1976. Moral Luck. *Proceedings of the Aristotelian Society, Supplementary Volumes* 50: 137-151.

Petersson, Björn 2004. The Second Mistake in Moral Mathematics is not about the Worth of Mere Participation. *Utilitas* 16: 288-315.

Pettit, Philip 2007. Responsibility Incorporated. *Ethics* 117: 171-201.

Rescher, Nicholas 1998. Collective Responsibility. *Journal of Social Philosophy* 29: 46-58.

Sadler, Brook Jenkins 2006. Shared Intentions and Shared Responsibility. *Midwest Studies In Philosophy* 30: 115-144.

Sher, George 2009. *Who Knew?* Oxford U. P.

Shockley, Kenneth 2007. "Programming Collective Control". *Journal of Social Philosophy* 38: 442-455.

Smiley, Marion 1992. *Moral Responsibility and the Boundaries of Community: Power and Accountability from a Pragmatic Point of View*. University of Chicago Press.

Strawson, Peter F. 1962. Freedom and Resentment. *Proceedings of the British Academy* 48: 1-25.

Sverdlik, Steven 1987. Collective Responsibility. *Philosophical Studies* 51: 61-76.

Sytsma, Justin; Livengood, Jonathan and Rose, David *ms*. Two Types of Typicality: Rethinking the Role of Statistical Typicality in Ordinary Causal Attributions. *Manuscript*

Tännsjö, Torbjörn 2007. The Myth of Innocence: On Collective Responsibility and Collective Punishment. *Philosophical Papers* 36: 295-314.

Wright, Richard W. 1988. Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof: Pruning the Bramble Bush by Clarifying the Concepts. *Iowa Law Review* 73: 1001–1077.

Zimmerman, Michael 1985. Sharing Responsibility. *American Philosophical Quarterly* 22: 115-122.

Zimmerman, Michael J. 2008. *Living with Uncertainty.* Cambridge U. P.