

Collective explanations, joint responsibility

Gunnar Björnsson

University of Gothenburg

Linköping University

1. Introduction

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 will happen. However, as the solvent from all three diffused throughout the lake over the next few days, its concentration became high enough to prevent micro-organisms in the lake from reproducing, thus leaving higher organisms without food and effectively wiping out all fish in the lake. The concentration of solvent exceeded the threshold for the micro-organisms 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 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; at least, this is how things have seemed to just about everyone that I have confronted with the case. Similarly, it seems that voters can be morally responsible for the outcome of a referendum, consumers for the practices of companies they patronize, and frequent fliers and drivers of SUVs for climate effects, even though they, as individuals, could not affect those outcomes, practices or effects. The question of this paper concerns the conditions for such joint responsibility for outcomes of collective actions.

The first answer people want to give is that although no individual could control the outcome, the group could, and the individuals share the responsibility of the group. If Alice, Bill and Cecil had been less reckless, the fish would still have been alive; their actions, taken together, made a difference to that outcome, and each contributed to that difference making. Even if that were correct—and we will see that it isn't quite right—it would still leave the question of *why* the individuals share the responsibility of the group. The fact that a group made a difference for an outcome does not *in general* imply that all the members are jointly responsible; especially not members who try their best to prevent the outcome.

Consider two examples where the board of a company votes for adopting a certain plan that is predicted to damage the environment and where, as a result, the environment is indeed harmed. One member of the board, Agnes, is against the plan because of these effects. In the first example, she votes against the plan. In the second, she votes for it, because she knows both that it will pass with or without her vote—the other members have no concern for the environment—and that voting for it is the only way in which she will have a chance to mitigate its effects (which she subsequently manages to do, to some extent). Agnes is part of the group in both cases and her action in the second is member of a set of actions—the acts of voting for the plan—that together made a difference to the environmental effects. In neither case, however, is it at all obvious that Agnes shares the moral responsibility for the negative effects with the other members of the board (although she might be *legally* responsible). This also means that if we think that *the company* or *its board* is morally responsible for the resulting environmental damage, that responsibility does not automatically *distribute* to each of the individuals that are involved in bringing it about.¹ By contrast, Alice, Bill and Cecil *all* seem morally implicated in the death of the fish. At the same time, it seems wrong to say of Alice that *she* is morally responsible for the death of the fish unless it is understood that she is responsible *together* with Bill and Cecil. After all, her action, considered on its own, made no difference to the outcome, and Bill and Cecil were involved in the very same way as Alice.

Furthermore, the *joint* character of Alice's, Bill's and Cecil's responsibility for the outcome is not well understood by the standard account of individual participation in responsibility for the outcome of collective actions. Such accounts invoke intentions to act

¹ Whether collectives can be responsible when *none* of their members are is an interesting question that has received quite some attention recently. For defences of the autonomy thesis, see (Arnold 2006, Pettit 2007, Tännsjö 2007, Copp 2007), for criticism see (Corlett 2001, Haji 2006, McKenna 2006, Seumas 2007)

together or intentions that are conditional upon the intentions of the others,¹ but none of the three polluters knew or had any reason to believe that the others were similarly reckless, and they had no way to discuss the issue or coordinate their actions. One might even want to deny that they formed a *group* or a *collective* in any ordinary sense—though I will sometimes talk about “collective action” when discussing a mereological sum of actions performed by different individuals.

What I will argue here is that the trio’s joint responsibility for the outcome is naturally understood given an account of moral responsibility that also explains a number of puzzling features of *individual* responsibility. According to this account, the *Explanation Account*, an agent is morally responsible for an outcome insofar as some relevant motivational quality of the agent is part of a significant explanation of the outcome. It is because it is quite natural to explain why the fish died with reference to, say, *the three polluters’ lack of concern for the environment*, that the trio is responsible for that environmental disaster in the small. Moreover, since the lack of concern and resulting action of *each* of the three is part of that explanation, each partakes of that responsibility. That is why they are *jointly* responsible for the outcome.

The paper is organized as follows: In the next section, I introduce some caveats and make some clarifying remarks to pinpoint the notion of moral responsibility that I am concerned with. In section three, I explain why a variety of ways of accounting for responsibility for outcomes of collective action fail to account for what goes on in this case and attempt an alternative diagnosis suggested by variations of the original case. In section four, I introduce the Explanation Account. In the final section, I show how the Explanation Account subsumes the diagnosis of section three and briefly discuss some implications for other putative cases of collective moral responsibility. One of the important consequences of the proposed analysis is that it shows just *how* responsibility for outcomes of collective action is a normative issue.

2. Outcome responsibility, luck, blameworthiness, wrongness

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 a number of other issues discussed under the heading of “responsibility”—in particular issues

¹ For accounts that take joint responsibility to involve representations of “joint action” in some way or other on part of members of the collective, see (Rescher 1998, Kutz 2000, Sadler 2006, Shockley 2007).

of legal liability, of moral or legal obligations to *take* responsibility for outcomes, of moral or legal obligations to *ensure* certain outcomes, and of conditions for being a responsible person, or decision procedure.¹

Furthermore, the concern here is with *outcome* responsibility rather than responsibility for intentions or 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 here are autonomous, in control of her own decisions and actions, capable of rational deliberation and suffering from no motivational maladies, and that, as a result, each is responsible for his or her own act or failure to act. The question here is on what grounds a number of individuals can be understood as *jointly responsible for outcomes*.

Some puzzles about outcome responsibility do not specifically concern outcomes of collective actions. In particular, many find it problematic to ascribe moral responsibility for outcomes that depend on factors outside the agent's control: such outcomes seem to be matters of luck. The stock example of outcome luck involves two drunk drivers. They are equally reckless but one runs over and kills a child who happens to run into the street just in front of the car. Had the driver been sober, she would have been able to break, but she fails because of the intoxication. The other driver gets home without incident, but would have been equally unable to break if that child had run into the street in front of her instead. In cases like these, it might seem deeply unfair that the lucky driver should be taken to be any less blameworthy (Enoch and Marmor 2007, e.g.). To the extent that moral responsibility implies desert of blame and punishment—and many make that connection—it would seem unjust to hold these people responsible to different degrees. This worry extends to cases of responsibility for outcomes of “collective” actions. Had Bill and Cecil disposed of the solvent in a safe way rather than poured it into the lake, the fish would have survived and Alice would not have been responsible for their death. Since Alice could not affect Bill's or Cecil's actions, it might seem unfair that they should affect her degree of blameworthiness.

¹ For example, although being responsible for a bad outcome in the relevant sense often brings with it a duty to respond to or compensate victims for that outcome, the fact that I have a duty to respond or compensate for an outcome does not imply that I am responsible for it in the relevant sense. My duty to respond might be grounded in the fact that victims can justifiably expect me to support the perpetrators (because they come from my country), and a duty to compensate might be grounded in the fact that I had unknowingly received stolen goods.

Since I cannot give a full defence of moral responsibility for outcomes in this context, I will follow most discussions of responsibility for outcomes of collective action and just assume that outcome responsibility is a defensible notion. Nevertheless, I will indicate some reasons that I am not too worried about the role of luck in determining outcome responsibility, reasons that might also give the reader a better understanding of the relevant notion of responsibility.

The first thing to hold in mind is that on this notion, responsibility does not directly imply desert of blame or punishment. For example, we might agree that someone is responsible for something—the toppling of a president elected under dubious circumstances, say—while disagreeing about whether or to what degree blame or praise is appropriate. Since responsibility doesn't imply blameworthiness, being more responsible for an event doesn't imply being more blameworthy. However, some deeper worries about luck will remain, for to deny that moral responsibility implies blameworthiness is not to deny other strong connections between moral responsibility, blame, praise or desert. In particular, that someone is morally responsible for an outcome will seem to many to be a *necessary* condition for taking her to deserve blame or praise for that outcome and perhaps also for taking the value of that outcome to affect the moral rightness or wrongness of her actions.¹ Some will no doubt think it objectionable enough that this precondition for certain moral assessments depends on matters outside the agent's control.

The second thing to notice is that when we are morally responsible and blameworthy for several bad things to different degrees, there is no simple way in which this adds up to an overall measure of blameworthiness. The notion of moral responsibility at stake here and the corresponding notion of blameworthiness are relational: degrees of responsibility are always degrees of responsibility *for something*. And this means that if one is fully responsible for drunk driving or for pouring solvent into the lake, the fact that one is also, to some extent, responsible for the death of a child or the fish in the lake does not automatically mean that one is *more* responsible or blameworthy than one would have been if the child had not run into the street or others had disposed of their solvent in a safe way. It just means that one is responsible or blameworthy for one more thing.

¹ But see Cushman (2008) for evidence that adults in general judge moral wrongness and permissibility based on the intentions of agents but assign blame based (also) on causal responsibility.

Readers who tend to be worried that moral luck undermines moral responsibility should hold these things in mind throughout the rest of the discussion.¹ Admittedly, moral responsibility understood in this way might not be able to play what such readers have taken to be its core role: to directly justify severe punishment and rewards in the spirit of *lex talionis* (cf Strawson 1994: 9-10). However, the strategy here will be to stay close to what seems to be common everyday intuitions about moral responsibility and to give an account of moral responsibility for outcomes of collective action that makes sense of these intuitions. If moral responsibility, so understood, is too weak to support a strongly retributivist ethics, so be it. Since this weaker notion seems ubiquitous in everyday moral thinking, we should at least try to understand it. (Moreover, as I will briefly indicate later on, properly understanding it also provides some fairly direct justification of our reliance on this notion.)

Finally, notice that being morally responsible for a bad outcome doesn't imply having acted wrongly, nor does being responsible for a good outcome imply having performed a good action. If nothing else, good and bad outcomes of an action might cancel out.

3. Carving responsibility for outcomes of collective action at its joints

The purpose of this section is to clarify what it is that governs intuitions about moral responsibility for collective outcomes in cases where there is no awareness of collective action or even of the possibility of collective action among members of the collective. To do this, I will present a series of variations of our initial example, thus hoping to isolate the aspects of the case that seem essential to our attribution of moral responsibility.

We have already noted that individual responsibility for an outcome of collective action does not rely on *individual difference making*. None of the individual acts of pollution made a difference to the effect we are concerned with here: the fish would have died as quickly even

¹ There are also reasons to allow matters of luck to affect desert, since everything we are and do—including decisions—seems to depend on matters of luck, or matters outside our control: on our genes, our experience and our opportunities to act and possibly on indeterministic processes. The even our decisions are matters of luck gets one of its standard expressions in Strawson's (1994) argument against the possibility of moral responsibility, but is also part of Nagel's famous (1976) argument for moral luck. Given this claim, to accept that people are responsible for their own decisions and actions is already to accept that morally responsibility isn't directly undermined by luck. Admittedly, this argument is undermined if we assign responsibility only because we are blind to the role of luck in our decisions. For an argument for the compatibility of luck and moral responsibility, see (Björnsson and Persson 2009a).

if there had been only two polluters. This is what makes responsibility for outcomes of *collective* action particularly puzzling. What is less often noticed is that it doesn't require *collective* difference making either, as shown by cases of what David Lewis (1986b) calls "causal preemption". Suppose that, unknown to the trio, a fourth person, Dave, were waiting in the background, continuously monitoring the levels of solvent in the lake with the intention to poison the micro-organisms should the level not rise high enough without his intervention. Then the fish would have met the same fate even if Alice, Bill and Cecil had disposed of their solvent in a safe way. But Dave's presence does not seem to change the fact that, given how events *actually* unfolded, the trio is to blame for the outcome.

Similar cases also show that the relevant connection between actions and outcome is not (just) one of being a necessary part of a sufficient condition for that outcome. Consider a case where Alice, Bill and Cecil switch places with Dave: Independently of each other, members of the trio set up gadgets that monitor the level of pollution in the lake, being set to dispose of solvent (the same amounts as in the original example) into the lake should the levels of pollution be low. However, Dave poured enough solvent into the lake to kill off the fish and prevent the gadgets from adding more solvent. Here, although the actions performed by the trio are pair-wise sufficient for the death of the fish and although the action of each member was a necessary part of such a pair, Alice, Bill and Cecil are not responsible for that outcome (though presumably blameworthy for their malevolent actions).

Furthermore, one can be responsible for an outcome in virtue of actions that are not necessary parts of a condition that is causally *insufficient* for the outcome, as cases of probabilistic causation show. Suppose that the solvent only affects the micro-organisms if it undergoes a chemical process that needs to be triggered by a certain sort of indeterministic but highly likely event. Then the actions of the three polluters would not be a sufficient condition for the death of the fish: they would not causally guarantee that outcome. Nevertheless, the polluters would still seem to be responsible for that outcome given that the process actually took place. (Moreover, this would be true even if Dave were waiting in the background, ready to start a process that would deterministically necessitate the death of the fish should the trio dispose of the solvent safely.)

In response to these difficulties, it is natural to think that what is required here is *causation*: the relevant set of actions needs to *cause* the outcome in order for the agents to be responsible for it. The difficulty of describing the relevant relation in terms of sufficient or necessary conditions is familiar from efforts to analyze causation, where various cases of overdetermination have provided seemingly endless problems for both counterfactual theories

of causation in the tradition of Lewis (1973) and accounts in the tradition of Mackie (1974) that understand causation in terms of non-redundant parts of nomically sufficient conditions (for discussion of difficulties, see e.g. Collins et al 2004, Björnsson 2007). If the relation required here is indeed that of causation, the difficulties we have seen are just what one could expect. It might also seem that understanding the relevant relation as one of causation has another important benefit: it seems to straightforwardly explain why the three polluters are all responsible for wiping out the fish. After all, solvent poured into the lake by all three was causally involved in preventing the micro-organisms from reproducing, thus contributing to the starvation of the fish.

Eventually I will indeed argue that the relevant relation between group action and outcome is one of causation, or of *explanation why*. 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 micro-organisms 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 some of 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 (Pettersson 2004). But that is not right either. Suppose that when the concentration of solvent reaches above that resulting from two polluters, the process by which the micro-organisms are prevented from reproducing is both slowed down and made more open to possible disturbances, thus making the outcome slightly less likely. Then it is true of each of the polluters that he or she actually lowered the probability that the fish would die and obstructed that process to some degree, given the actual contribution from the other two. Nevertheless, all three polluters would still seem to be jointly responsible for the death of the fish: it still died because of their actions.¹

¹ It is true, of course, that each polluter is increasing her *subjective* probability that the wildlife will be affected negatively. What we are trying to determine here, however, is what the required causal involvement in producing the outcome would have to be in order for the agent to be among those responsible for the outcome. Whatever the agent's beliefs are about what she is doing, she is not

We will return to the issue of what the relevant individual involvement has to be in order for an individual to be among those responsible for the outcome of collective action. Our immediate concern is to understand how the *collective* must relate to the outcome; the suggestion is that the relevant relation is one of causation. Though I take this suggestion to be basically correct, it is important to see that the relevant notion of causation is one that need not involve a relation of *bringing about* or *producing* but could be one of *letting things happen*, as a case involving omissions shows:

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, 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 to the extent that they would have been if they had actively pushed Hannah into the well), and they certainly seem responsible for the fact that Hannah wasn't saved. They could have saved her, but they did not.

In this case, unlike in *The Lake*, there is a sense in which none of the three was 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 action—or lack thereof—
→ responsible for the outcome unless her action is *actually* related to the outcome in the right way. In a case where she believed that the solvent was poisonous when it was in fact completely benign, she could be responsible for doing something she thought would have bad effects, but not, it seems, for any *actual* bad effects.

explains why Hannah wasn't saved. This is the notion of "explaining why" that seems relevant for our ordinary attributions moral responsibility in these cases.

The suggestion thus far is that in order for a group to be collectively responsible for an outcome, it has to be jointly involved in *the explanation* of that outcome.¹ But more needs to be said about the required sort of involvement. As we have already seen from the well case, the relevant involvement need not consist of any particular sort of positive *action*: perhaps Eric was sitting on a rock, Fiona climbing a tree, and George running across a meadow. Moreover, no decisions on part of members of the group need to be involved in the explanation. Perhaps none of the three 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. 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 had no concern for the environment and thus failed to react to the warning signs on the cans of solvent.

The last observation suggests that what needs to be involved in the explanation of the outcome isn't so much the behaviour of the agents involved as their *motivational states*: what they care and do not care about. Further variations on previous cases show this more clearly. Suppose that George did not hear Hannah's cries for help (perhaps he was walking on dry leaves next to a noisy brook). 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.

Judging from the variations of the two cases that we have considered, it seems that the two groups of people are responsible for the outcomes because the outcomes are explained by their motivation. The fish died because Alice, Bill and Cecil lacked appropriate concern for the environment; Hannah wasn't saved because Eric, Fiona and George lacked appropriate

¹ Since variations of the two cases indicate that probabilistic relations can ground explanatory judgments, this account seems to presuppose rejecting deductivism about explanations.

concern for their fellow human beings. Moreover, this correspondence between explanatory judgments and judgments of responsibility is robust across some further variations. First, in both *The Lake* and *The Well*, the outcome is explained by a *lack* of appropriate concern on part of the agents. Second, the agents in both cases seem to have acted *wrongly*. Neither of these features is necessary for the attribution of responsibility as long as the outcome is intuitively explained with reference to the agents' motivational states. First, Alice, Bill and Cecil would seem to be equally responsible for the outcome in a case where each sadistically wanted the fish to die rather than merely lacked appropriate concern. Correspondingly, it would still seem natural to explain why the fish died with reference to Alice's, Bill's and Cecil's motivational states: the fish would have died because of their sadism. Second, consider a case in which each member of the trio discovers and mends a leaking sewer out of concern for the environment, and where 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 (jointly) responsible for that outcome.

The variations considered thus far strongly suggest that what matters for attributions of joint moral responsibility for an outcome is that the motivational states of the individuals in the group are implicated in an explanation of that outcome. The question remains, though, whether we can expect this diagnosis to survive still further variations, and whether it generalizes to other cases of collective action. Also, 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 on the ground that (a) each individual is responsible for one wrongdoing—risking adverse environmental effects or not saving a fellow human being calling for help—and (b) what they risked actually took place because of these wrongdoings, taken collectively. Moreover, we have yet to explain why the *individuals* are jointly responsible for the outcomes, given this diagnosis. In the next two sections, I will provide reason to think that the account generalizes. As we shall see, the diagnosis conforms to a well-supported hypothesis about our concept of individual moral responsibility. Furthermore, we shall see how this concept applies straightforwardly to the agents in *The Lake* and *The Well*.

4. The Explanation Hypothesis and the Explanation Account

In two recent papers, 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.

We 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.¹

In order for our practices of holding people responsible to reliably promote or prevent certain outcomes in this way, they need to be targeted at types of motivational structures that are a) systematically tied to those outcomes and b) tend to be amenable to modification when targeted by these practices, on occasions when instances of the motivational structure type c) explains 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

¹ I am not denying 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 reformatory 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, or plausible evolutionary rationales for retributive tendencies.

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.¹

This provides motivation for what we call the “Explanation Hypothesis”, an empirical hypothesis about the conditions under which we hold people responsible. It says the following:

The Explanation Hypothesis: People take P to be morally responsible for E to the extent that they take E to be an outcome of a type O and take P to have a motivational structure S of type M such that GET, RR and ER hold:

General Explanatory Tendency (GET): Type M motivational structures are significant parts of a reasonably common sort of explanation of type O outcomes.

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): The case in question instantiates the right sort of general explanatory tendency: S is part of a significant explanation of E of the sort mentioned in GET.

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

¹ 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 (2009a) we explain how our particular way of spelling out this connection avoids some of the standard objections raised against such accounts.

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, the fact that we want something sufficiently often explains that we make it happen, guided by 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").¹ 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").²

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 it is asked why this happened. In answering that question, we could list a number of conditions, each of which might be a necessary part of complex sufficient condition for the

¹ 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. 6, Sher 2009).

² It is possible that GET should be restricted to these two broad kinds of explanation.

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.¹ 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. 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.

Because explanations are selective in these ways, what strikes us as a good explanation depends on our general expectations about the sort of event we seek to explain and the

¹ 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 probabilistic explanation where events might lack sufficient causes.)

situation in which it occurs. If we are concerned with an area where houses are frequently hit by lightning but tend to be well protected by lightning rods, we might want to explain why the house burned down by citing the absence of a lightning rod. Similarly, what we take to be a significant explanation might depend on *normative* expectations. Consider two cases where a concierge unsuspectingly lets two well-behaved gentlemen that turn out to be burglars into a building where they break into an apartment and take off with jewellery. In the first case, the concierge was supposed not to let any stranger unaccompanied by a tenant into the building; in the second case, there was no such rule. When we consider the first case and ask why the apartment was burglarized, we are likely to cite the fact that the concierge let the burglars into the building; not so when we ask about the second case.¹

Finally, our explanatory judgments are influenced by our explanatory interests. If we want to know why something happened in one case *but not in another*, we want to know some difference between the two cases that made the outcome more likely in the first than in the second. If we want to know why a given case had one outcome *rather than another*, we want to know what it was about that case that made one outcome more likely than the other. Both sorts of contrastive explanatory interests rule out a focus on otherwise important causal factors that are the same in both cases, or fail to make one outcome more probable than the other.²

When condition ER in the Explanation Hypothesis refers to a *significant* explanation, that means an explanation that satisfies our explanatory interests given our current normative and non-normative expectations or, differently put, an explanation that fits our *explanatory frame*. The selective and context-dependent 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 a wide variety of otherwise disparate phenomena. Some of these accounts are briefly indicated in the following list; the purpose here is merely to indicate the wide explanatory scope of the Explanation Hypothesis and fuller

¹ For empirical data illustrating the effect of normative expectations, see (Alicke 1992) and (Knobe and Fraser 2008).

² For one early contrastive account of explanation, see (van Fraassen 1980), see also (Schaffer 2005).

explanations along with explanations of other phenomena are given in (Björnsson and Persson 2009a, 2009b).

1. *External force and threats mitigate moral responsibility to various degrees.* The greater the threat or external force involved, the less difference in outcome is made by variation in the motivational structure.
2. *Lack of knowledge mitigates moral responsibility to various degrees.* When we have no idea that our actions will have or not have a certain outcome, caring more or less about that outcome will typically not affect our behaviour in ways compatible with GET.
3. *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.* For example, those who fail to intervene and stop a man from assaulting a woman might be responsible to some degree for the damage she suffered, but they seem decidedly less responsible than the perpetrator, other things being equal. We can focus on the causal contribution of the active culprit without focusing on those who allow things to happen, but we cannot see the explanatory role of the latter without taking into account that of the active culprit. The explanatory role of the latter will thus tend to grab our explanatory attention more strongly.
4. *Someone who takes initiative is more responsible than someone who tags along.* Often, taking initiative demands a more extraordinary motivation than tagging along. More importantly, however, the action of the person who tags along is explained by the action of the person who took the initiative, but not vice versa, making the latter a more natural centre of explanatory attention.
5. *Judgments of moral responsibility tend to be undermined by considerations suggesting that many of our decisions are a matter of luck.* Arguments from luck against moral responsibility are typically framed in terms of contrastive explanations that automatically shift motivational states into the explanatory background, producing an explanatory frame where they are less significant.
6. *Judgments of moral responsibility tend to be undermined by arguments claiming that, ultimately, our actions are the upshots of events over which we have no control.* Such arguments shift explanatory focus to events that are causally prior to our motivational states while glossing over the complexities involved in the how these prior events explain our motivation, thus shifting the explanatory focus away from the motivation of the agent.
7. *The felt conflict between determinism and moral responsibility is lessened when people consider concrete cases, especially concrete cases involving grave moral transgressions.*

Deterministic scenarios tend to shift explanatory focus away from the motivational states of those involved, but focus on the concrete details and especially details that provoke strong moral reactions counteract those effects.

8. *Judgments of moral responsibility tend to be undermined by reductionistic, mechanistic explanations of behaviour.* Such explanations are easily understood as competing with and thus undermining our everyday explanations in terms of agents' motivational structures.
9. *Most people ascribe a higher degree of responsibility for negative than positive side effects that the agent did not care about.* When the agent lets something bad happen, the outcome is often explained by her lack of concern: we have a normative expectation that she should care about these things and also an expectation that if she had cared enough, she would have acted so as to avoid the outcome. When she lets something good happen, however, this is not explained by her deviation from normative expectations: had she cared about this good, she would still have let it happen.

As I will argue in the next section, the Explanation Hypothesis also accounts for intuitions about collective responsibility in the cases we have looked at in this paper. All this gives it considerable support as an empirical hypothesis about how we form our *judgments* of moral responsibility. Moreover, the etiological, functional characterization of the concept of moral responsibility suggests a justification of our reliance on that concept: it plays a central role in regulating our practices of holding each other and ourselves responsible and so a central role in making and keeping us responsible.

In itself, though, the Explanation Hypothesis is not an account of *moral responsibility*. Because our explanatory judgments vary with our explanatory frames, the hypothesis predicts that we make different judgments of responsibility depending on our explanatory interests and non-normative or normative expectations. This means that two people with the same beliefs about what has happened in a particular case can diverge in their judgments: one will think that the agent is responsible for an outcome while another denies this. For example, a person who expect mothers to be strongly protective of their children will be much more likely to hold the mother responsible when her child falls and breaks an arm during some rough and tumble play, and a person with a strong expectation that values concerning sleep, food and exercise affect health will be more likely to hold a person responsible for her diseases and ailments. The Explanation Hypothesis says nothing about whether these judgments are *correct*, or whether judgments made with other expectations are more reliable. For that

reason, it can provide no *direct* support for the claim that the agents involved in *The Lake* and *The Well* are in fact jointly morally responsible for the outcomes.

However, we can formulate a related thesis about moral responsibility—call it “the Explanation Account”—if we qualify ER by demanding that the motivational structure should be a significant explanation of the outcome given the *relevant* explanatory frame: given *correct* normative expectations and *relevant* non-normative expectations and explanatory interests. The Explanation Account would then say that P is morally responsible for E when GET, RR and ER are satisfied given the relevant explanatory frame. Of course, without a substantial characterization of what the relevant explanatory frames are, the Explanation Account implies no determinate judgments of responsibility. But since our interest here is to map moral responsibility for outcomes of collective action by tracing how everyday intuitions about these things vary as cases are changed subtly, we can assume, pending arguments to the contrary, that the explanatory frames that produce these intuitions are the relevant ones.¹ Let us therefore see how the Explanation Account, given that assumption, accounts for the presence or absence of collective moral responsibility in *The Lake* and *The Well*.

5. The Explanation Account and responsibility for outcomes of collective action

The Explanation Account tells us both why the agents of *The Lake* are responsible for the death of the fish and why they are *jointly* responsible for that outcome. They are *responsible* for the outcome because the three conditions GET, RR and ER are satisfied, and they 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, (2) is clearly defective, for two reasons. First it brings attention to the fact that Alice's carelessness made no difference to the outcome

¹ In (Björnsson and Persson 2009a) 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.

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.¹ But (2) also seems defective 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.² 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 structure also satisfies RR: we have assumed that the individuals involved satisfy conditions needed for individual responsibility for decisions and action.

¹ 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 counterfactual dependence might undermine the claim that Alice's lack of care 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.

² Things would have been different if Cecil had less reason than the others to think that the solvent would be dangerous, say. In such a case, Cecil's motivational structure would provide a less significant explanans: she would have acted in the same way even if she had cared more about environmental impact (but not she had cared *much* more). Things would also be different if Alice and Bill are talking about what happened. It would make sense for Alice to say, "It is our fault that the fish died", because Cecil is absent from the conversation. (But it would also make sense for Bill to add, "... and Cecil's".)

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).

The treatment of *The Well* is almost exactly the same, although the defect of an explanation singling out one individual is more strongly marked. "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 (2) 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 micro-organisms whereas Eric's inaction made no definite difference at all.¹

What we have seen, then, is how the Explanation Account supports the diagnosis of joint responsibility provided in section 3. 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 a similar reason, we should hesitate before saying that typical intuitions about cases like *The Lake* result from confusedly attributing joint responsibility when all these cases provide are *non-distributive* collective responsibility *for an outcome* and *individual* responsibility *for decisions and actions*. For intuitions of joint responsibility, the argument suggests, relies on the same sort of reasoning as do intuitions about individual responsibility.

Before closing, let me briefly mention three consequences of the account. The first concerns the relation between joint responsibility and lack of individual control. What makes joint responsibility particularly interesting is that it can be ascribed in cases where no one individual could control the outcome. But having seen how joint responsibility for outcomes of collective actions falls under an account of moral responsibility that applies equally to individual responsibility for outcomes of individual actions, we might also consider the possibility that people are jointly rather than individually responsible in some cases where there *is* individual control. Think of a version of *The Well* where any one of Eric, Fiona and

¹ It might be worth noting how the Explanation Account 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, our 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 yields different explanatory frames: the former already implies that Hannah as in danger, thus relegating her initial fall into the well into the explanatory background.

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, and that explanans corresponds to the most natural assignment of responsibility, namely jointly, to all of them.

The second consequence of the Explanation Account concerns the relation between joint responsibility and non-distributive collective responsibility. The sort of joint responsibility assigned to the trios by the Explanation Account in this and either of the two original cases is *distributive* in the sense mentioned in the introduction: *each* agent is responsible for the outcome *together* with the other members of his or her trio. The motivational structure of each agent stands in the required relation to the outcome. This provides a tool for attributing quite different sorts and degrees of responsibility to different members of a group or organization that is causally responsible for an 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.

The final point concerns the relation between normative expectations and outcome responsibility, which makes the question of joint responsibility much more complex. Given high enough normative expectations that people should avoid working for or purchase the goods of corporations or nations that are responsible for certain outcomes, it will seem that great many people without direct involvement in a company's environmental policy, say, or in persecution of members of organized labour in South America or the enactment of Israeli policy on the West Bank are nevertheless responsible for these things. After all, if people had cared more and been more principled, these things would have been very different. But this 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?¹ The Explanation Account that it makes clear just how such questions are central to issues of collective responsibility, by being directly relevant for the identification of significant explanations.²

¹ Questions about strength of normative expectations are also of crucial importance for discussions of individual responsibility for personal health and questions of whether health care benefits should be somewhat sensitive to whether people are responsible for his or her health problems.

² One thing worth noting is that this normative component of judgments of moral responsibility pertains most strikingly to *outcome* responsibility. The explanatory connection between motivational structures and action is in itself striking enough to make the motivational structure a significant explanation of the action in normal cases.

Bibliography

- Alicke, M. D. 1992: "Culpable Causation," *Journal of Personality and Social Psychology* 63:3, pp. 368-78
- Arnold, Denis G 2006: "Corporate Moral Agency". *Midwest Studies In Philosophy* 30:1, pp. 279-91.
- Björnsson, Gunnar 2007: "How Effects Depend on Their Causes, Why Causal Transitivity Fails, and Why We Care about Causation". *Philosophical Studies* 133:3, pp. 349-90.
- Björnsson, Gunnar and Persson, Karl 2009a: "The Explanatory Component of Moral Responsibility", forthcoming in *Noûs*
- Björnsson, Gunnar and Persson, Karl "Judgments of Moral Responsibility 2009b: A Unified Account", *Society for Philosophy and Psychology*, 35th Annual Meeting 2009, available at <http://philsci-archive.pitt.edu/archive/00004633/>
- Collins, John; Hall, Ned; and Paul, L. A. (eds) 2004: *Causation and Counterfactuals*. The MIT Press.
- David, Copp 2007: "The Collective Moral Autonomy Thesis". *Journal of Social Philosophy* 38:3, pp. 369-88.
- Cushman, Fiery 2008: "Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment". *Cognition* 108, pp. 353-80.
- Enoch, David and Marmor, Andrei 2007: "The Case Against Moral Luck". *Law and Philosophy* 26:4, pp. 405-36.
- Haji, Ish 2006: "On the Ultimate Responsibility of Collectives". *Midwest Studies in Philosophy* 30:1, pp. 292-308.
- Knobe, J., Fraser, B. 2008: "Causal Judgment and Moral Judgment: Two Experiments," *Moral Psychology* Vol 2, ed. Sinnott-Armstrong, W., MIT Press, pp. 441-47.
- Kutz, Christopher 2000: *Complicity*, Cambridge U. P.

- Lewis, David 1973: "Causation". *Journal of Philosophy* 70, pp. 556–567. Reprinted in Lewis 1986a, pp. 159–172.
- 1986a: *Philosophical Papers*, Vol. II. Oxford U. P.
- 1986b: "Postscripts to "Causation". In Lewis 1986a, pp. 172–213.
- Mackie, John 1974: *The Cement of the Universe*. Clarendon Press.
- May, Larry 1990: "Collective Inaction and Shared Responsibility". *Nous* 24:2, pp. 269-77.
- Michael, Mckenna 2006: "Collective Responsibility and an Agent Meaning Theory". *Midwest Studies in Philosophy* 30:1, pp. 16-34.
- Nagel, Thomas 1976: "Moral Luck". *Proceedings of the Aristotelian Society, Supplementary Volumes* 50, pp. 137-51.
- Petersson, Björn 2004: "The Second Mistake in Moral Mathematics is not about the Worth of Mere Participation". *Utilitas* 16:3, pp. 288-315.
- Petersson, Björn 2008: "Collective Omissions and Responsibility". *Philosophical Papers* 37:2, pp. 243-61.
- Pettit, Philip 2007: "Responsibility Incorporated". *Ethics* 117, pp. 171-201.
- Rescher, Nicholas 1998: "Collective Responsibility". *Journal of Social Philosophy* 29:3, pp. 46-58.
- Sadler, Brook Jenkins 2006: "Shared Intentions and Shared Responsibility". *Midwest Studies In Philosophy* 30:1, pp. 115-44.
- Schaffer, Jonathan 2005: "Contrastive Causation". *The Philosophical Review* 114:3, pp. 297-328.
- Seumas, Miller 2007: "Against the Collective Moral Autonomy Thesis". *Journal of Social Philosophy* 38:3, pp. 389-409.
- Sher, George 2009: *Who Knew?*, Oxford U. P.
- Shockley, Kenneth 2007: "Programming Collective Control". *Journal of Social Philosophy* 38:3, pp. 442-55.
- Strawson, Galen 1994: "The impossibility of moral responsibility". *Philosophical Studies* 75:1, pp. 5-24.
- Strawson, Peter F. 1962: "Freedom and Resentment". *Proceedings of the British Academy* 48, pp. 1-25.
- Tännsjö, Torbjörn 2007: "The Myth of Innocence: On Collective Responsibility and Collective Punishment". *Philosophical Papers* 36:2, pp. 295-314.
- van Fraassen, Bas C. 1980: *The Scientific Image*, Oxford U. P.
- Zimmerman, Michael J. 2008: *Living with Uncertainty*, Cambridge U. P.