THE 35TH ANNUAL MEETING OF THE SOCIETY FOR PHILOSOPHY AND PSYCHOLOGY, JUNE 12-14, 2009, INDIANA UNIVERSITY, BLOOMINGTON, IN

Judgments of moral responsibility – a unified account

Gunnar Björnsson

Karl Persson

University of Gothenburg

Introduction

In recent years, experimental philosophers have tried to determine whether ordinary people take freedom of will and moral responsibility to be compatible or incompatible with determinism. The results show that folk intuitions about moral responsibility are prone to contradiction, being sensitive to a surprising variety of factors (Nichols 2004, Nahmias et. al. 2005, Nahmias et. al. 2007, Nichols and Knobe 2007). This has led philosophers to debate whether people can have any unified concept of moral responsibility (Nelkin 2007, Knobe et. al. 2007, Doris and Knobe *forthcoming*).

Elsewhere, we have presented an analysis of our everyday concept of moral responsibility that provides a unified explanation of paradigmatic cases of moral responsibility and accounts for the force of both typical excuses and the most influential skeptical arguments against moral responsibility or for incompatibilism. In this article, we suggest that it also explains the divergent and apparently incoherent set of intuitions revealed by some recent experiments. If our hypothesis is correct, the surprising variety of judgments stems from a unified concept of moral responsibility.

In what follows, we will briefly review the experimental results that are relevant for our purposes, sketch the model, and explain how the results are just what we could expect given that model.

GUNNAR.BJORNSSON@FILOSOFI.GU.SE KARL.PERSSON@FILOSOFI.GU.SE

Abstract vs. Concrete

Recent studies by Shaun Nichols and Joshua Knobe (2007) suggest that whether people take agents to be responsible for their actions in a deterministic scenario depends on whether these actions are described abstractly or concretely. Subjects were first presented with the deterministic scenario:

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

Subjects in the "abstract" condition were then asked whether, in Universe A, it is possible for a person to be fully morally responsible for his or her actions, whereas subjects in the "concrete" condition were asked whether a man called Bill intentionally kills his wife and children because he has become attracted to his secretary is fully responsible for killing his family. In the abstract condition, 86% percent answered *no*; in the concrete condition, 72% answered *yes* (Nichols & Knobe 2007:670).

The puzzle is to explain *why* people assign responsibility in the concrete but not in the abstract cases when the universe is portrayed the same way in both.[1]

[1] A number of philosophers have thought that different perspectives yield different judgments concerning moral responsibility. When we take the God's perspective on things, Daniel Dennett (2003: 92f) argues, responsibility seems undermined by inevitability. Sir Peter Strawson ([1965] 2003: 79f) argues that judgments of responsibility belongs to a participatory attitude toward people rather than an objective one. And on one interpretation, Kant (1785: 97-101) takes our autonomy to belong to the practical rather than the theoretical perspective (Cf. Korsgaard 1996: 162-7).

High affect vs. Low affect

Another result from Nichols and Knobe (2007) displays differences in folk intuitions about "high affect" and "low affect" cases presented against a deterministic background scenario. In the high affect case presented to the subjects, a man stalks and rapes a stranger; in the low affect case a man cheats on his taxes (ibid.: 675). When subjects were presented with the high affect case they where more inclined to ascribe full moral responsibility (64%) than when they were presented with the low affect scenario (23%) (ibid.: 676). The question, again, is why this is so? As Nichols and Knobe argue, it is plausible that emotions affect our judgments, but why?

Psychological vs. Mechanistic explanations

Nahmias et al (2007) argue that what keeps people from assigning responsibility is reductionism and mechanistic explanations of people's behavior, not determinism (Cf. Green and Cohen 2004).

They set up an experiment in which subjects in a reductionist condition were confronted with a deterministic scenario in which *neuroscientists* have discovered the *chemical reactions* and *neural processes in our brains* that completely cause our decisions and actions, and are themselves completely caused by events preceding our births. By contrast, subjects in the non-reductionist condition were confronted with a scenario in which *psychologists* had discovered the *thoughts, desires and plans in our minds* that completely cause our decisions and actions, and are themselves caused by events preceding our births.

When subjects were asked to what extent they agreed that people in this deterministic scenario should be held responsible for their actions, the level of agreement was significantly lower among subjects in the reductionist condition: 41% agreed at least somewhat vs. 87% in the non-reductionist condition (Nahmias et. al. 2007: 227).

Again, the question is why we get these results. It is conceivable that incompatibilist reactions to abstract scenarios that are neutral between reductionist and non-reductionist interpretations are partially due to subject's understanding them along reductionist lines. But if this is the case, it still leaves questions of why we give them a reductionist interpretation, *why* we take reductionism to undermine responsibility, and why affect significantly affects judgments.

Positive and negative side effects

A study by Knobe (2003) suggests that people are significantly more inclined to hold an agent responsible for bringing about bad side effects than for bringing about good side effects when the agent just doesn't care about these side effects. Subjects were presented with two scenarios involving a chairman of a board who takes no interest in environmental effects of his decisions. In one scenario, he knowingly allows a profitable program that will *harm* the environment; in the other he allows a profitable program that will *benefit* the environment. Subjects in the harm condition assigned a high degree of blameworthiness (4.8 on a 0 to 6 scale), whereas subjects in the benefit condition assigned a very low (1.4) degree of praiseworthiness (Knobe 2003: 193). Again, the question is how to explain the difference.

The explanatory component of moral responsibility

We will suggest that these four experimental results are all well explained by the fact that our everyday idea of an agent's being *morally* responsible for an outcome involves the idea that the outcome is *explained* by the agent's motivation in normal ways.

The central hypothesis – call it the "explanation hypothesis" – is the following:

> ***The Explanation Hypothesis***: People take P to be morally responsible for an outcome or action E to the extent that they take E to be *explained in normal ways by some motivational structure* of P that is of a kind that *can be modified by reactive attitudes*.[2]

[2] More precisely, the Explanation Hypothesis says that people take P to be morally responsible for E to the extent that they take GET, RR and ER to be satisfied:

> General Explanatory Tendency (GET): There is a reasonably common condition C such that motivational structures of type M explain outcomes (actions, events) of type O given C while motivational structures of type M' explain outcomes of type O' given C, where M and M' as well as O and O' are mutually exclusive.
>
> Reactive Response-ability (RR): Generally speaking, whether people exemplify M or M' depends on whether they are subject to being held responsible for realizing or not preventing O or O'.
>
> Explanatory Responsibility (ER): P has a motivational structure, S, of type M; E is of type O; C holds; and S is part of a significant explanation of E (in the normal way that M-motivation explains O-outcomes given C).

The hypothesis is not meant to explain *everything* that people say about moral responsibility, but we suggest that it is responsible for most typical intuitions about moral responsibility, including the reactions that are driving the philosophical debate.

There are two kinds of reason to accept the explanation hypothesis. Elsewhere, we have argued that one can expect something like the explanation hypothesis to be true given the role judgments of moral responsibility serve in governing our reactive attitudes. For these attitudes to serve their social function of controlling and shaping motivational structures and promoting and prevent various events, they need to be directed towards the sort of motivational structures that (a) are causally responsible for these events in systematic ways and (b) respond to reactive attitudes in the appropriate way. Since they are directed by our concept of moral responsibility, it is reasonable to expect that concept to keep track of just these conditions. And this, of course, is just what the explanation hypothesis says that our judgments of moral responsibility do.

This functionalist story is enough to make the explanation hypothesis interesting, but what is more important is its capacity to explain a number of general features of ordinary thinking about moral responsibility. For example, it seems plausible that our sensitivity to whether outcomes are explained by kinds of motivation that are *modifiable by reactive attitudes* is what leads people to assign diminished moral responsibility for compulsive behavior or for behavior performed under extreme emotional stress.[3] Moreover, we generally take people to be responsible for things that would not have taken place if they had been motivated by different things or to a different degree: for outcomes that are explained in normal ways by their desires for such outcomes, and to be responsible for bad outcomes that are explained in normal ways by their lack of concern for such outcomes. And we generally do not take people to be responsible for things that they are physically forced to do, or for outcomes that no one could possibly have predicted: in such cases, the agent's motivation is typically explanatorily irrelevant for the outcome.

[3] Here it is important to keep in mind that the explanation hypothesis is concerned with motivational structures of certain *types*, not with the instantiation of motivational structures in a particular individual. If a reckless driver dies as he crashes into another car, we hold that driver responsible even though death obviously prevents any further changes to the motivational structure.

The explanation hypothesis has much more interesting consequences, however, having to do with the fact that the notion of what explains E is *selective* in a certain way. Ordinarily, when we are looking for the causal explanation of some event or condition, E, we are not trying to assemble any or all conditions or events that can be said to make a causal contribution to the occurrence of that event; we are not asking for a complete and maximally detailed description of its causal origins, or a complete explanation of why it came about. We are trying to identify some especially *interesting* explanatory condition, X.

Typically, X only provides a causal explanation of E given a number of further conditions, C, which we might call the *supporting conditions* of X's explaining E. Nevertheless, as we think that X explains E, our focus is on X and E, while C is part of the cognitive background of our thought; cognitively, X and E are treated as variables, while C is treated as a constant. Other parts of a maximally detailed description of how E came about are neither seen as explaining E, nor as supporting conditions for X's explaining E. Some of these are the conditions or events that explain X, which are ignored as our focus lies on the explanatory connection between X and E. Others are causal upshots of X and intermediary causal steps between X and E. Like E, these are understood as dependent on X, but get less focus than E when we focus on the thought that X explains E.

One of the factors that determine whether X is an *interesting* cause of E is whether X is more remarkable, surprising or out of the ordinary than the background conditions (cf. Hart and Honoré 1985: 33-44). When the smoke detector sounds its alarm, a complete causal explanation of the event will include various facts about the wiring of the detector, the fact that it has a good battery, and the presence of smoke. However, given that we expect the detector to be in good working order, what we would think of as *explaining the alarm* is the presence of smoke. If we had expected the presence of smoke but not that of the battery, we would have thought of the latter condition as what explained the alarm.

The interest-relativity of everyday explanatory judgments is well known, but has surprising explanatory power when our judgments of moral responsibility are understood as selective and interest relative in this way. Elsewhere, we have argued that it explains patterns of everyday excuses, such as the presence of threats, lack of control, conformity to socially expected pattern ("just did my job") and lack of active involvement ("I was just an innocent bystander"). Here, we will see how it explains our puzzling set of experimental results.

Explaining the results

We begin with the question of why people are less prone to ascribe responsibility to actions set in a deterministic scenario when these actions are abstractly characterized than when they are concrete. According to the explanation hypothesis, this must be because characterizations affect the explanatory judgments that we are ready to make. In this case, two factors lead away from the kind of explanatory interests that govern everyday judgments of responsibility and typically lead us to assign responsibility to intended outcomes of actions: the abstracting away from any particulars of the case, and the focus on conditions that explain the motivational structure of the agent without themselves being explained by that structure.

These two aspects combine to make motivational structures of agents explanatorily insignificant. First, when a cause of an event is itself caused by another event, our explanatory judgments will typically focus on the prior cause, unless the prior cause forces us to make assumptions that were not already taken for granted. For illustration, suppose that we knew the following:

> Sam arrived late for a meeting. One driver had been using her mobile phone, while another was having an argument with his wife; both were slow to react to changes in traffic and bumped into each other. One thing led to another, and a number of cars crashed hard into each other, blocking three out of four lanes for over an hour. Sam spent over 40 minutes behind slow-moving cars that were stuck behind other slow-moving cars, … , making their way past the site of the accident.

In briefly answering the question why Sam arrived late, we would presumably pick out from the chain of events the fact that a road accident had blocked all but one lane. This would provide an explanation of both the slow traffic and the fact that Sam was late, without invoking assumptions that were not already part of common background assumptions; that accidents blocking most lanes cause traffic congestion and late arrivals goes without saying. We would probably not just say that Sam got stuck behind slow-moving cars, as that would seem to imply that there was no significant explanation of why the cars were moving slowly. Nor would we be likely to mention the causes of the accident. Since arguments and telephone calls in cars typically do *not* lead to accidents, these prior causes would only make the late arrival intelligible if supplied by further assumptions making it clear that an accident followed and that cars were blocking the lanes. Although more would be explained, the explanation would be much less straightforward.

In this case, and when we normally assign responsibility, an explanatory regress is blocked by the fact that invoking prior causes would unduly complicate the explanation. But in the case of an abstractly characterized action set in a deterministic scenario, we are led to abstract away from the particulars and think in terms of events caused by prior events, thus eradicating differences in perceived explanatory complexity between an agent's motivation and what might be enormously complex sets of prior conditions causally responsible for that motivation. For this reason, nothing counteracts our tendency to prefer prior causes. When we ask for what explains a certain action, nothing picks out the motivation of the agent as particularly interesting. Given the explanation hypothesis, that means that agents will not seem to be responsible for their actions.

When an agent's actions are described in more detail, everyday explanatory interests will instead become more salient, and our tendency to focus on prior causes less pronounced. This, we suggest, is what explains the contrast in judgments of responsibility between abstract and concrete conditions in the experiments by Nichols and Knobe (2007). But it also provides a straightforward explanation of why subjects more readily ascribe responsibility in cases of intuitively more serious moral transgressions or "high affect" cases than they do in "low affect" cases. The reason is simply that more serious cases are more prone to grab our attention and thus more prone to prevent us from ignoring the particulars of the case and looking further back in the causal chain.

The explanation hypothesis provides a similar account of why a reductive, neurological, deterministic scenario undermines judgments of explanation much more strongly than a non-reductive, psychological scenario. Again, the reason is that the former case is more likely to change our explanatory interests. In the reductive scenario, we are not only lead to adopt a new set of explanatory categories (chemical and neurological processes) but also, perhaps, to discard our everyday explanations of human action, whereas the explanatory categories postulated in the non-reductive scenario coincide with those employed in everyday explanatory discourse. Given the explanation hypothesis, the results provided by Nahmias et. al. (2007) are to be expected.

Finally, consider the question of why people are significantly more inclined to hold an agent responsible for bringing about bad side effects than for bringing about good side effects when the agent just doesn't care about these side effects. The reason, we suggest, is simply that in the case with the bad side effects but not in the case with the good side effects, we take the agent's lack of motivation to *explain* the outcome: in the first scenario, we want to say that the environment was damaged because the chair of the board didn't care; in the second, we simply do not want to say that the environment was helped because of some aspects of the chair's motivation.

Final remarks

In this paper, we have argued that the explanation hypothesis accounts for the puzzling and seemingly incoherent variety of intuitions about moral responsibility that has been revealed by recent work in experimental philosophy. Much more can be said, of course, about possible alternative explanations, about whether the explanation hypothesis can explain other aspects of our thinking about moral responsibility. Moreover, the very fact that the explanation hypothesis might be able to explain a seemingly incoherent variety of intuitions leaves the question of which intuitions we *should* rely on when considering whether moral responsibility is compatible with determinism. But these questions are better left for another occasion.

References

Dennett, D. (2003) *Freedom evolves*. Viking Penguin

Doris, J. and Knobe, J. (Forthcoming) *Strawsonian Variations: Folk Morality and the Search for a Unified Theory* in *The Handbook of Moral Psychology*, ed. John Doris. Oxford U P

Green, J.; Cohen, J. (2004) For the law, neuroscience changes nothing and everything. *Phil. Trans. R. Soc. Lond. B*, 359, pp.1775–1785

Hart, M. and Honoré. (1985) *Causation in the law.* Oxford U P

Kant, I. (1785) *Grundlegung zur Metaphysik der Sitten*

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

Knobe, J.; Doris, J.; Woolfolk, R. (2007) Variantism about responsibility. *Philosophical perspectives* 21, 183-214

Korsgaard, C. M. (1996) *Creating the kingdom of ends*. Cambridge U P

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2005) Surveying Freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology 18*

Nahmias, E.; Coates, J.; Kvaran. T. (2007) Free will, moral responsibility, and mechanism: experiments on folk intuitions. *Midwest studies in Philosophy* XXXI

Nichols, S. (2004) The folk psychology of free will: fits and starts. *Mind & Language*, Vol. 19, No. 5, pp. 473-503

Nichols, S.; Knobe, J. (2007) Moral responsibility and determinism: the cognitive science of folk intuitions, *Noûs* 41:4, 663-685