An Anatomy of Moral Responsibility

Matthew Braham Martin van Hees

June 6, 2009

Abstract Can we attribute moral responsibility in cases of joint action, that is, when states of affairs are engendered by the actions of two or more individuals? Thompson (1980) has argued that it is often very difficult to do so and has christened this 'the problem of many hands'. More recently, Pettit (2007) claims to have found a 'problem of no hands', in which it is possible to say that a collective agent is morally responsible for bringing about an outcome but none of its constituent members are. In this paper we argue that such 'in principle' conclusions and existential pointers to shortfalls in individual responsibility are too quick. They arise primarily for a methodological reason: the appropriate formal analysis has not been undertaken. To trace the connection between individual agency and outcomes that arise from a complex of interactions we make use of a game theoretic framework. We show how individual responsibility can be described in such a framework, and examine the formal conditions under which responsibility assignments can be made.

1. Introduction

Judgements of moral responsibility for actual states of affairs – who, in general, is to be blamed, who is to pick up the bill or who gets the credits – are intimately connected to the relationship between an agent's behaviour and the state of affairs in question. Roughly speaking, a theory of moral responsibility selects a subset of agents to be subjected to moral criticism (praise or blame) and/or social sanctions (reward or punishment) on the basis of an assessment of the individual actions that are considered to have a significant connection with the state of affairs. In many situations, political ones in particular, many different agents contribute in a variety of ways to a particular decision or policy so that ascertaining who is to be held responsible will often be difficult. This has become known as 'the problem of many hands' (Thompson 1980).

There are many generic examples which demonstrate the difficulties of assigning moral responsibilities in cases of complex joint action.[1] A particularly prominent case is that of

[1] See Bovens (1998) for a collection of real-world cases.

‘Tragedy of the Commons’ (Hardin 1968). The ‘tragedy’ is one in which a group of independent and economically interested and active agents derive benefits from a subtractable resource but drive that resource to complete depletion as a result of aggregate use. In Hardin’s parable, herders share a common pasture of which no herder or group of herders has an enforceable property right, i.e.,the pasture is a ‘common-pool resource’. Assuming certain economic, demographic, and ecological conditions to hold, under every contingency it is always disadvantageous for each herder to reduce the size of their herd. The upshot is that the productivity of the pasture will inevitably be driven to complete depletion.

The situation is, of course, an n-person Prisoner’s Dilemma which has become a standard game theoretic model of a plethora of social and political dynamics. The choice never to reduce herd size is the dominant strategy as far as profit maximizing goes. The Prisoner’s Dilemma, and collective action problems in general, has two very important features which make the ascription of moral responsibility particularly troublesome in such circumstances. First, no individual has direct control over the outcome in the form of actions that are necessary or sufficient causal conditions for it. Second, even if we could unravel the problems of causal ascription, moral responsibility traditionally requires something more. For moral criticism, there must, in Feinberg’s (1970: 196) terminology, be a form of ‘causal relevance’ in the sense that the behaviours that were causal factors are also ‘faulty’ in some manner. What ‘faulty’ behaviour means in general is undeniably a contentious issue; and this is particularly so in the Tragedy of the Commons as we have described it, because it is hard to claim that there is anything intrinsically faulty about bringing livestock to pasture. The Tragedy of the Commons points to a very thorny axiological issue: the value of an act or object may depend on what else there happens to be. It is, thus, a very open question as to who is morally responsible for the overgrazed pastures and the eventual destitution of the herders.

The question is of course not only of theoretical interest but also of utmost practical importance. Given the connection between (negative and positive) sanctions and judgements about moral responsibility, we want to ensure that our assessments about who is, and is not, to be held responsible are justified. The connection also has obvious policy implications. Pinning moral responsibility is a ‘weak’ incentive device in the sense that approvals and disapprovals of our behaviour, even if only by – in Mill’s language – ‘the reproaches of our own conscience’ lead to attitudinal changes and changes in our future behaviour and thus to better or worse states of affairs. To illustrate the high stakes, suppose we are debating the issue of whether moral obligations regarding global warming and the harm that ensues in ecologically sensitive environments devolve upon individual people, given that global warming can be traced back to the countless individual decisions on a day-to-day basis. As many have done, we can model this situation as a Tragedy of the Commons.[2] In this case the ‘commons’ is the climate and in one version suggested by Johnson (2003) and Sinnott-Armstrong (2005), the actions are

[2] This is the basic position of the now well-known Stern Review (Stern 2007). Other problems of justice such as world poverty can also be modelled as a Tragedy of the Commons.

whether or not to take a spin in a gas-guzzeling sport utility vehicle (SUV) on a beautiful Sunday afternoon; while the outcomes are the manifold effects, which include the personal enjoyment ('the wind in my hair', 'the get up and go', 'the views', the 'excitement') as well as the increased level of carbon dioxide emissions, climate change, and eventual destruction of ecologically sensitive habitats. If we can pin moral responsibility on Sunday pleasure riders then, at least *prima facie*, an effective policy for alleviation of global warming should include educative measures that effect attitudinal changes at the individual level. If, as both Johnson and Sinnott-Armstrong believe, Sunday pleasure riders cannot be held morally responsible for global warning, then any such measures would be unjust in the sense that they place undeserved moral burdens on individuals. In such cases resources for mitigating climate change are better used for other purposes.

One way of pinning down the morally responsible in complex interactions is to adopt the concept of *collective responsibility*. This is the idea that a group *per se* can be charged with moral responsibility and then let the moral responsibility descend to the constituent members of the group by dint of some set of criteria such as 'mere membership', 'shared values', 'being a beneficiary' etc.[3] The problem with this approach is that, firstly, it is beset with taxing metaphysical and normative quandaries: the hypostatization of groups is anything but straightforward;[4] and the problem of membership criteria is that, as Lewis (1948) observed a fair while ago, this form of responsibility may implicate each one of us in one another's actions so that praise and blame will then fall on us without discrimination. Furthermore, in some situations it may be very difficult to relate the collective outcome to individual responsibilities. Indeed, in a recent and important paper, Pettit (2007) argued that there exists circumstances in which a collective agent is responsible for the realization of some state of affairs although none of the individuals constituting the collective agent is. He labels this the 'problem of no hands' and considers it to be a major moral and practical problem because if such responsibility gaps exist, the members of the group have an incentive to organize their activities in a self-serving way and at the same time be able to avoid any moral responsibility in the event of a harm occurring (Pettit 2007: 197).

The objective of this paper is ambitious. We set out to construct the necessary and sufficient conditions for assigning moral responsibility to individuals in complex decision situations. The analysis can be called formal in two meanings of the word. First, we focus on what we believe to be the anatomy or structure of the concept of moral responsibility, rather than that we present a substantive theory of responsibility of our own. Secondly, it is formal in the sense that we use a particular format: that of game theory. The advantage of doing so is that we can adumbrate a set of conditions that more or less define an algorithm for cutting through the thicket of concepts and examples that characterize the analysis of moral

[3] The list of possible criteria is very long. For an overview, see the collection of papers in May (1992). A recent discussion can be found in Miller (2007).

[4] For a treatment of the metaphysical problems see List and Pettit (2006) and Pettit (2007).

responsibility in general and cases of joint actions in particular.

The main argument of the paper is constructive: the general framework that we provide – the anatomy – can be seen as a result in itself. A set of more general technical results are provided in a companion paper (Braham and van Hees 2009). In section 2 we set out some conceptual preliminaries concerning moral responsibility. The conception that we will propose in a very terse way is that of responsiveness to 'reasonable demands' which is a generalization of ideas that can be found in, among others, Fischer and Ravizza (1998) and Scanlon (1998). In section 3 we introduce and expound upon the basic game theoretic concepts that we use. Drawing on (Braham and van Hees 2008), section 4 introduces and summarizes the conception of causal contribution that constitutes one of the components of our account. Here we will take on board the so-called 'NESS test', which is an increasing accepted and general account of singular (or actual) causation, and provide a game-theoretic formulation of it. In section 5 we operationalize the key component of our theory of moral responsibility: the *avoidance potential*; and in section 6 we set out our formal theory in full. In section 7 we apply our formal theory to the analysis of the Tragedy of the Commons and Pettit's 'no hands' claim. We close with some concluding remarks in section 8.

2. Preliminaries

As there are manifold conceptions of the concept of moral responsibility we need to begin by defining some terms of art and bracketing out certain issues. To avoid any misunderstanding, we will not work through the thorny metaphysical issues that surround the topic (such as whether or not free will is a necessary condition), nor provide a detailed defence of the concepts that we use – that is not the purpose of this paper. Rather, we will merely state these concepts, taking them on board as primitives, and explicate any principle distinctions that are called for in much the same way we state the assumptions and initial definitions when we develop and analyse a formal model in economics or political science.

As a start, two important restrictions need to be stated. First, the variety of moral responsibility that is at issue in this analysis is *retrospective*: it is about the allocation of moral responsibility for the actual realization of some outcome state of affairs. It is about what has happened, such as a pasture being over-grazed, the climate altered, or Jones killed. We will not say anything at this stage about *prospective* moral responsibility which refers to obligations toward the realization of some future outcome or state of affairs.[5] Second, we are concerned with determining the conditions for judging whether certain forms of conduct makes an individual an apposite target of negative or positive moral appraisal and not with judgements about virtue or vice – often called aretaic judgements – that constitute a form

[5] The distinction between outcomes and states of affairs will be clarified in section 3. We use the terms 'restrospective' and 'prospective' broadly and set aside some of the subtle issues that beset these terms, for which see Zimmerman (1988: 1–2).

of character assessment. To use Watson's (2004) helpful terminology, we focus on 'responsibility as accountability' and not on 'responsibility as attributability'.[6] Responsibility as *attributibility* is to make the agent 'the author' of her conduct. By doing so one can demand from the agent an explanation of the ends and values in virtue of which she choose to behave. A (positive or negative) appraisal of those actions then serves as the basis of appraisal of the agent's character, values, or goals.

When we hold a person *accountable* we are not primarily appraising her character or the goals that underlie her actions, but are using our evaluation of the actions as a basis of particular sanctions (praise, blame, or punishment).[7] Clearly, judgements of accountability often presuppose attributability, at least in the weak sense of the action in question being volitional. The converse need not be true.[8] We may condemn a person for doing the things she does and yet not hold her accountable – say because her behaviour is none of our business (e.g. a person deliberately squandering her talents) or because she did not have the option of choosing differently (e.g. when sadistic behaviour can be more or less traced back to physical and psychological abuse in childhood as in the infamous case of Robert Harris (Cartwright (2006))).

In particular, we assume that a focus on accountability implies that the actions of a person held responsible made a difference. In our approach this comes down to demanding that a person's actions were a *causal* factor in the emergence of the state of affairs. That is, we subscribe to the so-called entailment thesis: responsibility implies causation (Driver 2008; Sartorio 2004). We thereby adopt a theory of *event-causation*. Both causes and consequences are taken to be events, and consist of the enactment of strategies and the realization of states of affairs respectively. The relevant events often consist of actions but may, on our account, also consist of omissions. Thus Jones taking his cows to the pasture is an event (action), his not taking the cows to the pasture (omission) is also an event, as is the pasture having exceeded its carrying capacity (the realization of a state of affairs). The metaphysics of these events will simply be assumed to be consistent with our conception of event causation.[9] Not any kind of behaviour is a relevant event for moral responsibility. We take attributability as a necessary condition for accountability and thus take an 'action' or 'omission' to be more than just an arbitrary piece of behaviour (such as a bodily movement) or the absence of one; it must have a volitional quality in the sense of flowing from some form of autonomous intentionality. That is, the individuals in our analysis are assumed to have a capacity to reflect upon their

[6] See also Scanlon's (1998) distinction between substantive and attributive responsibility. Attributive responsibility has also been called *agent-responsibility* (Vallentyne 2008).

[7] We are mindful of the fact that the relationship between conduct and character is a complex one and has a bearing on moral appraisal of conduct but we set it aside here for reasons of tractability. Note that this means we will also set aside such issues as 'weakness of the will' (*akrasia*).

[8] However, as Feinberg (1968) for instance points out, it may be rational from a standpoint of incentives – particularly in law – for breaking the link between responsibility and authorship of an outcome.

[9] We are acutely aware that this is a somewhat heroic assumption but tractability requires that we set aside a discussion of this issue.

beliefs and desires and exercise some form of control over what they do, or refrain from doing: they are 'reasons responsive' and planning agents.[10] Phrased differently, the behaviours that we have in mind as inputs into a judgement of moral responsibility must be 'doings' and not mere 'undertakings'. Hence, we make use of a very weak concept of attributability. We will not defend this here but simply assume it is indeed sufficient for our purposes.

An important issue when allocating any form of responsibility concerns the information that the agents in question have. If an agent knew, could have known, or should have known, that a certain action of hers could lead to a particular outcome, then we will assess her moral responsibility for the outcome differently than if she could not have possibly known, or need not have known, the outcome would have resulted. Again, we are going to simplify the analysis by focusing on situations in which individuals have *complete information.* That is, they know the actions and omissions available to them and to others, and also know which outcomes follow from which particular combinations of actions. The only information they do not have pertains to what the other agents will do. Since the consequences of an agent's actions are partly dependent on those other actions, the individuals are often ignorant about the exact consequences of their action. The justification for restricting the settings to complete information is two-fold. Firstly, we see our analysis as a first step towards a full-fledged account of responsibility. Once we have established the conditions for allocating moral responsibility in an idealized context we can turn to more realistic contexts of incomplete and asymmetric information. Secondly, the lack of information that we do consider and which consists of the uncertainty about the actions performed by the others, is for the particular purposes of this paper of high importance. After all, the many hands problem arises in situations in which the outcome depends on the interactions of many individuals.

Finally, we need to turn briefly to Strawson's (1962) *social practice* account of moral responsibility. Strawson ascribes moral responsibility to a person if she is the justified target of reactive attitudes (such as blame). In doing so, he argued against what we may call the classical paradigm in which responsibility is analyzed. This paradigm consists of, first, the delineation of the sufficient and necessary conditions for ascribing responsibility; next, the application of the resulting notion to a specific situation; and, finally, the specification of the appropriate normative reaction (the reactive attitude). As we do seem to follow that paradigm here, a remark by way of justification is in order. First of all, we are interested in a particular type of moral responsibility, i.e. accountability, and it is quite possible that reactive sentiments may be targeted at a person for reasons other than being involved in the realization of the outcome or for being the author of it. The involvement may, for instance, be vicarious rather than contributory. Secondly, not every reactive attitude is always justified (Wallace 1994). People may overreact: they can be angry for the wrong reasons or handout praise

[10] The individuals can be said to be 'Davidsonian agents' after Davidson (1971). Note that this form of agency entails certain mental capacities such as the ability to detect a means–end coherence between a course of action and a specific outcome or state of affairs.

to the wrong person, etc. Since we do not want to build a theory of moral responsibility on reactive attitudes that cannot be justified, we need a criterion for establishing which actions are appropriate targets of blame and praise and which are not. Such a criterion can be provided for by the sufficient and necessary conditions for assigning responsibility if the formulation of those conditions is motivated by the question whether reactive attitudes like praise or blame would be justified. This is indeed our approach and it therefore is not an instance of the classical paradigm but rather lies somewhere in between the classic top down (from moral responsibility to reactive attitudes) and the Strawsonian bottom up (from reactive attitudes to moral responsibility) approach.

With these restrictions and caveats in mind, we now lay down what can be considered the canonical conditions for retrospective moral responsibility, conditions that we will adopt.[11] It contains three distinct components which we call the 'tri-partite analysis' of moral responsibility:

Definition 2.1 *An individual i is morally responsible for some outcome x if and only if:*

1. *i's conduct was the result of an appropriately autonomous choice;*
2. *i's conduct was a causal factor for the outcome;*
3. *i had reasonable opportunity to do otherwise.*

These are not independent conditions. For instance, on our account of causation a person's conduct can only be causal factor if the person had an alternative course of action available to him (though not necessarily a reasonable one). Similarly, whether a choice was made autonomously is not independent from the question of whether the agent had a reasonable alternative.

3. The Formal Framework

The elementary concept that we need throughout is that of a *game*. A game G is an n+4-tuple $(N, X, S_1, \ldots, S_n, \pi, R^n)$, where (1) N (with cardinality n) is a finite set of *agents* (players), (2) X a finite set of *outcomes*, (3) for each $i \in N$, S_i is a finite set S_i of possible *strategies*, (4) π is an an *outcome function* from the set of all strategy combinations $\times_{i \in N} S_i$, or plays, onto X (π being onto X means each element of X is an outcome in at least one play, with a play being denoted by a strategy profile $s_N = (s_1, \ldots, s_n)$), and (5) R^n is a preference profile, that is, an n-tuple of preference orderings (one for each individual) over X.

Very generally, a game specifies the 'rules of the game' such that it precisely defines who can do what, when, and to what effect, as well as the individuals' preferences regarding the

[11] These conditions are discussed in one form or another in, for instance: Cane (2002), Feinberg (1970), May (1992), Nagel (1979), Pettit (2007), Scanlon (1998), Vallentyne (2008), Zimmerman (1988).

various outcomes. We use the information provided for by a game as the basis of responsibility assignments. Obviously, the specification of the relevant game is, therefore, not an innocuous choice. To see this, consider Beebee's (2004) 'Queen of England Problem'. Suppose i promised to water j's plant but for whatever reason i failed to do so with the consequence that the plant died. To model this situation it is reasonable to focus on a game in which i and j are the only players. Now suppose we derive the judgement that j caused the plant's death. It may well be the case that on the theory of causation yielding that judgement we would also have to infer that the Queen of England's failure to water the plant is a cause of the plant's death: the reasoning that lead to the judgement that i's omission is a causal factor may also apply to the Queen's omission.[12] However, the Queen is not assumed to be part of the game and thus how a game is defined already selects the relevant individuals whose behaviour is to be appraised.

In the context of Mackie's (1965; 1974) theory of causation, which we adopt, the ingredients of the analysis are what he calls the 'causal field'. In the context of causation, this presupposes we know which part of the causal chain that led to some action is relevant for the assessment of a person's causal contribution and which part is not. In the 'Queen of England Problem' j's action is taken to be part of the causal field, whereas the Queen's omission is not. Given that we are developing an account of moral responsibility, we could say that the game not only specifies the 'causal field' but also the 'moral field'; considerations that are not included in the game are left out of the scope of the analysis.[13]

To give another example of the relevance of which game we take to analyze responsibility, suppose we want to determine which members of a voting committee are to be held responsible for a decision they made. If we restrict our attention only to the actual decision, then the game will only describe the situation at the time of the actual vote: the given set of players, the voting rule, the various outcomes to which the different combinations of votes would have led, and the actual outcome. Given this restricted information, we may come to the conclusion that some member of the committee can be exonerated because, say, she voted against the outcome that eventuated. However, if we extend the game by incorporating information about, for instance, how the committee came into being in the first place, that committee member may well be morally responsible after all – say because without his agreeing to join the committee the vote may not have occurred.

Taking some specific game as the starting point of the analysis, that is, presupposing a 'moral field', does not make the analysis arbitrary. First of all, we often are interested in assigning moral responsibility in well-defined specific contexts, such as when we want to

[12] On the account of causation that we present in the next section this is indeed the case. After all, had i watered the plant it would have lived; and had the Queen of England watered the plant it would have lived. On this problem see also McGrath (2005); Sartorio (2007).

[13] Mackie actually draws on Hart and Honoré's (1959) seminal analysis of causation in the law. They say that a causal ascription picks out unusual behaviour among prior contextual considerations. That is, a causal ascription is already normatively laden.

know who is responsible for a specific committee decision insofar as that responsibility can be traced back to the decsion-making process *within* the committee. In these cases, the research question determines, at least partly, the moral field. Secondly, the relevance of the moral field points out that assignments of moral responsibility are partly determined by our prior moral expectations: there is, we assume, no normatively neutral way of arriving at responsibility judgements. This means the choice of the game should be carefully justified but it does not mean such justification is impossible.[14]

Our next remark concerns the set X. The elements of this set are social states, that is, detailed descriptions of the world. While the elements of X are called *outcomes*, the subsets $A \subseteq X$ are called *states of affairs*. A state of affairs describes one or more aspects of an outcome. Thus, whereas x may, for instance, be the outcome in which Bob is elected to be the new leader of the labour union, A may represent the state of affairs in which a male person is elected to be the new leader and B the set of outcomes in which no new leader is chosen, etc. States of affairs are described extentionally. Hence, we have $x \in A$ and $A \cap B = \emptyset$.[15] The distinction between states of affairs and outcome is important. A person may be responsible for an outcome without being responsible for all aspects of it (I may be responsible for Bob being elected but I need not be responsible for the fact that a man is elected, as there may not have been a female candidate on the ballot); and, conversely, one may be responsible for a state of affairs without being responsible for the outcome (I may be responsible for the election of a male candidate – say because I vetoed all female candidates – without being responsible for the election of Bob).

A *strategy* is a course of action available to the agent. Generally speaking a strategy is a 'bundle' or 'complex of actions'. For instance, the strategy 'shoot' consists of all those physical events that result in a 'shooting' such as picking up the gun, aiming, curling a finger around the trigger, etc. A game treats 'omissions' as strategies. A strategy of 'not shooting', for instance, is any set of actions which does contain not certain actions which are necessary for 'shooting'. Omissions are not therefore, in Lewis's (2004) language 'absences' that cannot be considered causal relata 'by reason of their nonexistence'.

Finally, the outcome function π assigns to each combination of strategies a particular outcome. The function is completely abstract: it may be determined by empirically ascertainable 'laws of nature', such as the strength that is required for, say lifting a fallen tree that has trapped a person; or it may be a matter of social law and convention, such as a decision rule which determines who has the authority to do so certain things. In any case it can be either a statement of certain regularities of either 'natural' or 'conventional' generation.

The players, their strategies, the outcomes and the outcome function together form what

[14] The problems here are inherent inherent to any description. See Sen's (1980) insightful analysis of the choice problems of description.

[15] To avoid cumbersome notation, we shall always drop the set brackets when a state of affairs describes *all* aspects of an outcome, i.e when it is a singleton set. Hence, in such cases we denote both the outcome and the state of affairs by x, rather than by x and $\{x\}$, respectively.

is usually called a *game form.* A game form in itself does not give us information about what the agents will do. For that purpose game theorists also specify the *preferences* of the individuals and use a *solution concept*, which is an assumption about how rational individuals will act given the rules of the game and their preferences. The existence of a preference profile expresses the assumption that agents are able to order their wants, desires, needs, and interests in a consistent way. A solution concept, designated by Γ, specifies how rational individuals will play the game and consists of a subset of plays. We shall sometimes refer to these plays as 'equilibria' and the corresponding outcomes as 'equilibrium outcomes' though we do not commit ourselves to a particular solution concept (such as Nash Equilibrium or its refinements). At the most elementary level, a solution concept imposes a form of causal regularity on the players in that they have characteristic ways of adjusting themselves to background conditions.

What is relevant at this point is to note that in our constructive exercise we adopt 'rationality' in a very general and weak form. We need only assume that the players seek to make choices that are advantageous to them, whatever that may mean (it can include altruistic preferences) and that they are able to reason in an instrumental way and make consistent choices. It is in this sense that we say the choices made by rational individuals are the result of an appropriately autonomous choice. Clearly, this is a very weak definition of autonomous choice. It does not say anything about the nature of the preferences that an agent is maximizing. Suppose someone is threatened with the infliction of some terrible retaliation if he does not adopt a particular course of action. To give in to the threat is a simple exercise of utility maximization, and on our account would thus satisfy 'the autonomous choice' demand. However, as our formal apparatus will highlight, moral responsibility will not be assigned to a person in such a circumstance as it violates our third requirement in the tripartite analysis, viz. the availability of reasonable alternatives.

4. Making a Causal Contribution

In order to ascribe moral responsibility to an individual for some state of affairs, we must parse precisely her causal connection to that state of affairs. Before beginning two remarks are necessary. Firstly, although we use the term 'causation' or ' causal efficacy', we are aware that on occasion it is more precise to use the expressions 'contributory effect', 'conventional causality', or 'conventional generation' because the term 'causality' generally refers to the connection between two events that are related by some 'regularity' or 'laws of nature'. The cases that we will will focus on are governed by legal norms and social conventions and not merely by laws of nature as such. For instance, that a house burns down following the outbreak of a fire of a certain size follows from the regularities that we call 'laws of nature'; that a particular policy is implemented following the agreement of a given set of people follows from legal rules and conventions. That is, the states of affairs in our examples may not be nomically related to their 'causes'.[16] To adopt this other terminology would, however, burden the discussion without adding anything. In fact, we need not do so since we will primarily be concerned with causation for *particular* or *singular* events (or 'cause in fact' in legal terminology).[17] The conception of a cause that we will adopt thereby, but not defend, is that of *difference-making*. In Lewis's (1973: 557) paraphrase of Hume: 'We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have been.'

We will assume, as is generally the case in legal theory, and especially in tort law, that a cause is a relation of dependency to be understood in terms of necessary or sufficient conditions (Honoré 1995). In particular, we assume that it is a form of dependence that subordinates a criterion of necessity to that of sufficiency and replaces the idea of identifying some event as 'the cause' to that of a 'causally relevant factor'. This conception ascribes C causal status for E if it satisfies the following criterion known as the NESS test (Wright 1988: 1020):

Definition 4.1 *There is a set of sufficient conditions for E such that: (1) C is a member of the set; (2) all elements of the set obtain; (3) C is necessary for the sufficiency of the set.*

In words: C is a causal condition for E if C is a *n*ecessary *e*lement of a *s*ufficient *s*et of conditions for E. Or, somewhat more precisely, C is part of a set of conditions together sufficient for E and is necessary for that set of conditions to be sufficient for E.[18]

For our purposes it is already helpful to note that the NESS test easily accounts for cases of causal overdetermination. The reason is that an event is attributed causal status even if, due to the presence of other actually or hypothetically sufficient sets, it was not necessary in the circumstances for the result. To see how this works, suppose three individuals are walking in the woods and they come across an injured jogger trapped under a fallen tree trunk. It takes at least two to lift the trunk and rescue the jogger but as it happens all three do the lifting. There are three possible sets of actions that are minimally sufficient for the rescue and

[16] This distinction is discussed in more detail in Kramer (2003: 280).

[17] For the reader unfamilar with the literature on causation, the term 'singular causation' comes from the 'singular–general' distinction of types of expressions. For propositions about causation, we say that 'Mack's drinking of a gallon of wine was a cause of his drunkenness' is a statement of singular causation. In contrast, 'drinking a gallon of wine causes drunkenness' is a general statement and implies a covering law.

[18] The NESS test was actually first stated in Hart and Honoré (1959) and can be traced back to J.S. Mill. The NESS test was also formulated by Mackie (1965, 1974), in terms of INUS conditions: 'an *i*nsufficient but *n*ecessary part of a condition which is itself *u*nnecessary but *s*ufficient for the result'. Note that Mackie's (1965) original formulation was more restrictive than the NESS test as discussed in Wright (1988) because it contained a condition that ruled out causal overdetermination (condition 4), which he later dropped (Mackie 1974). For a critique of the NESS test as an account of causality, see (Cane 2002). Halpern and Pearl (2005) provide a fully fledged formal structure of the NESS test that takes into account some of the problems that the NESS test faces. We do not, however, need the apparatus here. Finally, we note that the NESS test as formulated here is often restricted by imposing the additional requirement that the sufficient sets must be *minimal* in the sense that no proper subset of the events is itself sufficient for the outcome in question (see Mackie (1965, 1974), Wright (1988), and Halpern and Pearl (2005)). We, will, however not use this strengthened version because it suffers a number of paradoxical problems which are discussed in Braham and van Hees (2008).

each rescuer belongs to at least one (in fact two) of these. Consequently each of the rescuers' actions can be attributed causal status.[19]

To formulate this notion of causality in game-theoretic terms, we need to introduce some additional notation. First, for all $T \subseteq N$, we call an element s_T of $\Pi_{i \in T} S_i$ a T-event: it describes the event of the members of T performing the actions described by s_T (if $T = \emptyset$ we may call s_T a non-event). Given an event s_T, s_i denotes the strategy of $i \in T$, for event s'_T, s'_i is the element played by $i \in T$ in s'_T, etc. Furthermore, we write (s_T, s_{N-T}) to denote the play of G which consists of the combination of the (mutually exclusive) events s_T and s_{N-T}. We let $\pi(s_T)$ denote the set of outcomes that can result from the event s_T: $\pi(s_T) = \{\pi(s_T, s_{N-T}) \mid s_{N-T} \in \times_{i \notin T} S_i\}$.[20] Note that $s_\emptyset = X$.

Definition 4.2 *A T-event s_T is a* sufficient condition *for $A \subseteq X$ if and only if $\pi(s_T) \subseteq A$.*

For any s_U and s_T, call s_U a *subevent* of s_T if $U \subseteq T$ if each member of U adopts the same strategy in s_U as in s_T. Abusing notation, we shall write $s_U \subseteq s_T$ to indicate that s_U is a subevent of s_T. Similarly, we say that s_U is a *proper* subevent, and write $s_U \subsetneq s_T$ if U is a proper subset of T.

Definition 4.3 *Given a play s_N, individual i makes a causal contribution to A (her actions were a causal factor) if, and only if, there is a subplay s_T of s_N such that*

1. *s_T is a sufficient condition for A,*

2. *the subevent $s_{T-\{i\}}$ is not a sufficient condition for A.*

Note that the second clause entails that the individual had an alternative strategy which could have led to a different outcome. If $s_{T-\{i\}}$ is not sufficient for A, then by definition there is a strategy s'_i for i and a combination of strategies s'_{N-T} for the players outside of T such that $\pi(s'_i, s_{T-\{i\}}, s'_{N-T}) \notin A$. Thus, on our account to be a causal factor entails the availability of an alternative course of actions which *might* have avoided the realization of the state of affairs. However, as we shall see in the next section this kind of alternative possibilities entailed by our account of causality is too weak to constitute moral responsibility.

Before proceeding, we need to acknowledge a limitation that arises from our focus on games in normal form. The NESS-test has problems dealing with some types of strategies in normal forms games. In particular, strategies comprising conditional actions create problems. To see this consider this well-known example from the causation literature. Assassin$_1$ and Assassin$_2$ are planning to kill Victim. They both want Victim killed and they know they both want this. Assassin$_1$ has the opportunity to poison Victim in the morning. If he does so, the poison will take its effect in the course of the day and Victim will die in the late

[19] More examples are discussed in Braham and van Hees (2008).

[20] As explained in Section 3, we do not want to use different notations for the *outcome* x and the *state of affairs* consisting of the set of which x is the only element and we shall therefore write $\pi(s_T) = x$ rather than $\pi(s_T) = \{x\}$.

afternoon. Assassin$_2$'s plan is to shoot Victim, for which he will have the opportunity to do so in the early afternoon. Suppose they know of each other they are contemplating these options, and suppose also that Assassin$_2$ at the time at which he could shoot Victim will know if Assassin$_1$ had poisoned Victim (say because of some small physical signs). Further, it is certain that Victim will die if either Assassin implements his plan. If we model this as a game in normal form Assassin$_1$ has two strategies, 'Poison' and 'Don't poison'. Assassin$_2$ has two unconditional strategies ('Shoot regardless of whether Victim has been poisoned' and 'Don't shoot regardless of whether Victim has been poisoned') and two conditional ones ('Only shoot if Victim has been poisoned' and 'Only shoot if Victim has not been poisoned'). Now suppose Assassin$_1$ had indeed poisoned Victim, and that Assassin$_2$ had chosen the strategy of only shooting if Victim had not been poisoned. Clearly, Assassin$_1$ is causally effective for Victim's death. However, if we were to apply the NESS-test we also have to conclude that Assassin$_2$ is causally effective for Victim's death. After all, the strategy of Assassin$_2$ is a sufficient condition for the death of Victim. This is clearly counterintuitive. Whereas Assassin$_2$ had every intention to kill Victim if Assassin 1 would not do so, he did not actually shoot Victim.

The problem arises because the NESS-test only generates a convincing causal ascription of a person's actions if it is applied to the actions the players *actually* performed. It yields counterintuitive results if it is applied to conditional actions. For decision situations such as these, we should apply the NESS-test to the underlying games in extensive form, rather than to the game in normal form. We will, however, not go into this issue in this paper.

5. The Avoidance Potential

The demand that the person could have done otherwise has two components. The first is *feasibility.* Since the outcome of a game depends on the actions of various individuals, we should – when examining the possible consequences the adoption of an alternative strategy has – take account of these actions. In fact, we should not only consider what could have happened given the strategies others *actually* adopted, but, as we have already intimated, also given the actions they *might have* adopted.

Consider Joan who wants to go with two of her friends to the wedding of a mutual friend. She calls them and asks what their plans are. The first friend tells her that he has tickets for a Bob Dylan concert and that he rather go there than go to the wedding with his friends. The second informs her that she would like to go to the wedding but only if the three of them go together. Otherwise, she would prefer to go to the cinema. Given these preferences, Joan decides to join her second friend and goes to the cinema. As it later turned out, the wedding was a fantastic party and a one-off reunion of far flung friends. Joan rightly blames herself for having missed the wedding. However, is she also responsible for not having been at the wedding *in the company of her friends*? She made a causal contribution to the non-realization of that state of affairs as well: her strategy of going to the cinema is a minimally sufficient condition for it. Yet, it is counterintuitive to say that Joan is morally responsible for it. To assign responsibility the alternative outcome should be 'feasible' given the agent's information about the state of the world and given the justified expectations she has about what the others will do. To pinpoint this feasibility requirement, we examine for each strategy what we call its *avoidance potential.* This is the kernel of the contribution of this paper. Call a contingency s_{N-i} feasible if it is part of some equilibrium play.

Definition 5.1 *Given a solution concept* Γ*, the avoidance potential of a strategy* $s_i \in S_i$ *for a state of affairs* A*, denoted* $\rho_i(s_i, A)$*, equals the number of plays* (s_i, s_{N-i}) *in which*

1. $s_{N-\{i\}}$ *is a feasible contingency, i.e., is part of some equilibrium play*
2. i *is not causally effective for the realization of* A.

The avoidance potential should not be confused with the idea of control over outcomes (such as is found in van Inwagen's (1978) 'Principle of Possible Prevention'). When calculating the avoidance potential of a strategy we do not examine whether the strategy may yield *a state of affairs different from* A but whether it *fails to be a causal factor* for the realization of A. This distinction is important because our definition of avoidance potential merely cuts the causal connection between the strategy and the outcome, while the stronger definition would establish a causal link to an alternative outcome.

To get to grips with the avoidance potential, consider the following variation of the game with two assassins, call it Two Snipers. The two assassins are in place as snipers and will have to make at the same point in time a decision to shoot or not. Suppose Sniper 2 has a strict preference for shooting Victim himself – even when Sniper 1 also shoots. As it happens both do indeed shoot. Given his strict preference, and since he is assumed to be a utility-maximizer, Sniper 2 will shoot Victim. Is Sniper 1 responsible for the death of Victim? He is, just like Sniper 2, causally effective for it by way of the NESS-test. The fact that, given Sniper's 2 preferences, he did not have a course of action open to him which would lead to a different outcome does not vitiate his responsibility. The crucial point is that he could have avoided being a causal factor for the realization for x. Now, while this will not come as a surprise to those familiar with the literature surrounding Frankfurt's (1969; 1971) seminal contributions on moral responsibility, the explanation we provide for this is entirely new: it has nothing to do with intentions and second-order volitions etc., but with the formal structure of the avoidance potential.[21] However, it is also clear that 'control' is not entirely banished from moral responsibility. The concept of avoiding a causal contribution to the outcome implies that the person had control over the *state of affairs* in which the outcome occurred. Furthermore, and it is important to note – though it is a trivial point – that it is only with the alternative actions of the others that we take feasibility to be relevant, a

[21] It will, however, be a challenge to those who believe that control of outcomes is a necessary condition for moral responsibility. See, for instance, Morriss (1987: 39) for an example of this view.

person's own alternative actions need not concern us. To illustrate, suppose Sniper 1 also had a strictly dominant strategy for killing Victim. In this case there was only one feasible outcome, to wit, the one in which he kills Victim. Clearly, we do not let him off the hook; and it would be a strange theory of moral responsibility that would.

The notion of the avoidance potential adds precision to the very basic idea that a person is morally responsible for the realization of some state of affairs A if he was causally effective for A and if he could have adopted a strategy which had a higher avoidance potential for A. More precision, however, is needed. The second aspect of the idea of a having an alternative option is that the option forms a *reasonable* opportunity, i.e. that it is reasonable to demand that the person performed it.

Suppose a car jacker threatens a driver with his life if he does not yield the car to him which, as he tells him, will be used for a bank robbery. In the face of the threat, the driver surrenders his car and, as it later turns out, the car is indeed used as the getaway car in a bank robbery in which innocent people are injured. Even though the avoidance potential for the bank robbery is lower for the strategy of not giving the man his car, we do not want to hold the car owner responsible for the bank robbery – given that his life was at risk, it would not have been reasonable to demand from him to try to resist the car jacker. Similarly, and less grim, we want to be able to distinguish between Sunday pleasure rides in a gas-guzzling sport utility vehicle – 'don't be a spoilsport, I'm having such fun' – and the contribution that comes about from using that vehicle for the same distance to drive someone to hospital – 'Don't be ridiculous, my wife was in labour'.

Thus the avoidance potential needs to be refined in a specific way. The alternative strategies that make up the potential should 'eligible' or 'acceptable' by some standard. We shall use the term *eligibility*, which we take from the freedom literature. The notion, which was introduced by Benn and Weinstein (1971) and later discussed by Day (1977), Jones and Sugden (1982), and Sugden (1998), is that when we evaluate a person's freedom to do something we generally have to make some restrictions about what these 'things' or 'doings' are. Assuming, for our purposes, a conception of freedom in which freedom consists of the possibility for an agent to perform some action or actions of various kinds we may be faced with all sorts of possibilities. There are opportunities, however, that do not appear relevant to the assessment of our freedom. Cutting off our ears is an example. The reason being, in Benn and Weinstein's opinion, is that it 'is not the sort of thing anyone, in a standard range of conditions, would reasonably do, i.e.,"no one in his senses would think of doing such a things" (even though some people have, in fact, done it)' (Benn and Weinstein 1971: 195).

The nub of the matter is that, in this view, the presence of an *ineligible* opportunity to do something cannot be said to be a relevant freedom and thus to affect our freedom.[22] In the same way as an ineligible opportunity should not be considered as affecting our freedom

[22] This does not deny that it is in fact a freedom in the sense of purely unconstrained opportunity and that it may of value to have such freedoms. For this, see Carter (1999).

in some substantial way, the availability of an ineligible strategy should not be considered as affecting our responsibility. A demand that we should have behaved otherwise than we did, in the sense that we should have performed some alternative action, is reasonable if that action in question is eligible. A person is to be excused for the realization of some bad outcome, or does not deserve praise for a good one, if avoidance of it was only feasible by the adoption of an ineligible strategy.

This general idea has found a variety of expressions. Fischer and Ravizza (1998) say that a person is morally responsible for some outcome if they can recognise that there are sufficient reasons to have done otherwise.[23] Wallace (1994: 7) says that moral responsibility requires 'the power to grasp moral reasons, and the power to control one's behaviour in the light of such reasons'. As explained in the introduction, it is not our objective to take a particular stance or defend a full-fledged normative theory about when a person can be said to have reasonable opportunity to do otherwise. We presuppose the existence of some such theory and give the conception formal content as follows: for each individual i, eligibility is a mapping $\mathcal{E}_i$ which assigns to each contingency $s_{N-\{i\}}$ a subset of the individual's strategies which, given those actions of the others, are eligible. Note that eligibility is taken to be a context dependent notion. To illustrate, if the $N - \{i\}$ players perform actions that would result in serious wrongdoing unless i tells a lie, then the act of lying may well be eligible even though in other circumstances (i.e.,if the actions of other agents could not result in wrongdoing) it may not be. As we are not interested in this paper in fleshing out a normative theory (the substance of 'reasonable alternatives'), but only in investigating structural features – the anatomy – of a particular form of moral responsibility in games, we will refrain from any further exploration of the nature of these eligibility functions.[24]

6. Necessary and Sufficient Conditions for Moral Responsibility

We have now reached the stage where it is possible to give a crisp account of our method for assigning moral responsibility. It consists of two formal conditions that are necessary and sufficient for selecting the set of individuals who are legitimate objects of all that goes along with the attribute 'morally responsible': praise or blame, negative or positive sanctions, and criticism. Call a 'responsibility game' a triple $(G, \Gamma, \mathcal{E})$, where G is a game, Γ a solution concept, and $\mathcal{E}$ a set of individual eligibility functions (one for each individual).

Definition 6.1 *In a play $(s_i, s_{N-\{i\}})$ of a responsibility game $(G, \Gamma, \mathcal{E})$ an individual i is morally responsible for a state of affairs A if and only if:*

1. *s_i was a causal factor for the realization of A;*

[23] Fischer and Ravizza call this 'reasons-responsiveness' and it comes in a *weak* and *strong* form. A similar idea is suggested by Nozick (1981: 317–362).

[24] See however van Hees (2008) for a recent analysis of such functions in the context of assessing the value of our freedom of choice.

2. there is some s'_i *in* $\mathcal{E}_i(s_{N-\{i\}})$ *for which* $\rho_i(s'_i, A) > \rho_i(s_i, A)$*.*

The definition is general in that it accounts for the ascription of moral responsibility in one- or n-player normal form games of complete information. Except for the conditional strategies discussed in section 4 the analysis is suitable for a very broad class of collective action: most social dilemmas in which assignments of moral responsibility are clearly relevant (collective action problems) can be modelled in the normal form, as are most types of voting.

The definition is a 'result' in itself because it provides a clear set of variables and a relation between these variables for discussing moral responsibility. That is, the definition informs us about the data and base-relations that we must account for before we can legitimately pass judgement on a person's behaviour with respect to some outcome. Many of the classic examples that are discussed in the modern literature on moral responsibility do not actually take into account the full range of variables that we have introduced in our definition. So, for instance, none of the examples that are discussed in, say, Fischer and Ravizza (1998) or Sartorio (2007), explicitly mention preferences and solution concepts; and while, these are mentioned, for instance in Johnson (2003) and Sen (1974), these authors do not refer to explicitly to *all* the components of the avoidance potential: their discussions are basically restricted to the eligibility restriction. Similarly, while Goldman's (1999) responsibility-based solution to the 'paradox of voting' – it is not rational to do so – is close to our definition because it is based on the NESS-test, it does not take into account feasibility restrictions.[25] It is also worth noting that the definition is silent about socially optimal states of affairs. No necessary burden is placed on a player to choose the socially optimal outcome (by some standard) and then hold them as morally responsible for it not occurring if they happen to have chosen a strategy leading to a sub-optimal outcome.

To avoid any misunderstanding, although the second clause contains a quantifier, we do not impute *amounts* of responsibility to the players. At no point have we taken a stand on, or discussed, whether moral responsibility can be quantified or distributed in a particular way. At best we can say that, when the conditions are fulfilled, an agent is said to bear *some* responsibility for the realization of the state of affairs in question.[26] Quantification is a task that reaches beyond the scope of this paper, although admittedly the notion of a strategy's avoidance potential does provide a very natural basis for such quantifications.[27]

Finally, we note that the definition has both descriptive and normative content. On the one hand, we claim that *any* definition of retrospective moral responsibility that connects

[25] Goldman's solution in fact is a special case of Definition 6.1 because for the type of voting games that he considers feasibility will be redundant.

[26] Although we often make judgements of the form i bears more moral responsibility for x than j, or that i and j share moral responsibility for x in some proportion, not everyone agrees such statements are meaningful. See Zimmerman (1985) for an argument against the notion that moral responsibility can be expressed in relative shares of a fixed sum.

[27] In Braham and van Hees (2009) we demonstrate how such quantifications are possible for purely causal contributions.

outcomes to individuals will take this form in one way or another. That is, any ascription of this form of moral responsibility will refer to a game, the solution concepts (standards of behaviour), a causal condition, an eligibility function, and the avoidance potential. On the other hand, the definition can be taken as normative: *any* definition of retrospective moral responsibility *ought* to take this form, i.e. our constructive exercise demonstrates the proper way to make such ascriptions.

7. Application

In this section we apply our account of moral responsibility to two prominent examples that we mentioned in the introduction: the Tragedy of the Commons and Pettit's (2007) 'the problem of no hands'.

First, consider the Tragedy of the Commons or, more generally, any collective action situation that can be modelled in this framework (global warming, depletion of the sea, production of public goods, volunteering). As economically active agents, individual tries to maximize her utility as best as possible but the result of this individually rational behaviour is a state of affairs A which they all regret as it makes them all worse off than some other state of affairs that they could have reached if they had coordinated their strategy choices. To analyze whom is to be held responsible we could decide to present the game in detail, state the solution concept, define the appropriate eligibility mappings and subsequently check whether our conditions for assigning responsibility are satisfied. However, we need not do so. For our purposes it suffices to restrict ourselves to two characteristics that all such situations have in common. First, each individual i has a strategy s_i^* which is 'strictly dominant', that is, for any contingency $s_{N-\{i\}}$ and any alternative strategy s_i of his, she strictly prefers $\pi(s_i^*, s_{N-\{i\}})$ to $\pi(s_i, s_{N-\{i\}})$. This describes the fact that for each individual there is a uniquely optimizing strategy. Secondly, each individual i has at least one strategy s_i' such that in the play $(s_i', s_{N-\{i\}}^*)$: (a) i would not be causally effective for A, and (b) s_i' is eligible. Thus each individual has at least one eligible course of action s_i' which is not causally effective for the outcome. The strategy s_i' is usually some sort of restriction on the side of the player, say a restriction of the number of cows a farmer brings to the commons, a reduction of one's 'carbon footprint' activities, etc. Note it is only assumed that each individual has some such 'restricting strategy', they may in fact have more. It is also important to point out that not all types of restrictions need be eligible: if not sending any of his cows to the commons means that a farmer and his family will starve to death for lack of income, then it is not reasonable to demand that he do so. We only assume that *some* strategy of restricting oneself is eligible. Now assume that the equilibrium concept in place (i.e. that which is adopted by the players) is Nash equilibrium. In the game there is a unique one, the play in which all individuals choose their dominant strategy, and we assume that it is indeed how the game is actually played: none of the individuals restricts herself, as a result of which the common pool resources are depleted.

Precisely, then, our proposition is that given Definition 6.1, it is not only true that *at least one* individual can be ascribed morally responsibility but *all* individuals can be so ascribed. To prove this, we first show that some individual is causally effective for the depletion. Let T be a subset of N such that the play s^*_T leads to an outcome in A, that is, to depletion. Clearly there is at least one such T, viz. N itself. Now let T' be a smallest subset of T for which this is true, that is, T' is a group of farmers whose unrestricted grazing policy is 'minimally sufficient' for depletion of the lands. Then, by the NESS-test each member of T' has indeed played a causal role in depleting the land. By the second characteristic, each member of T' has one eligible strategy s'_i, the avoidance potential of which is 1: there is only one feasible contingency, $s^*_{N-\{i\}}$ and i is not causally effective for A in the play $(s'_i, s^*N - \{i\})$. Since, as we have shown, i is so in the actual play, the avoidance potential of his actual strategy is 0. Clearly, this holds for every member of T' and we can infer that all the members of T' are morally responsible for the realization of A – they had reasonable opportunity to do otherwise. Because T' cannot be empty, we conclude that at least one individual is indeed morally responsible. If we now make the extra assumption – one which is often made only implicitly – that each of the strategies actually played is indeed a member of at least some minimal sufficient condition, i.e. each individual made a causal contribution, then we can conclude that *each* individual is to be held responsible.

We now turn to a recent argument put forward by (Pettit 2007) that we mentioned in the opening section of this paper. He argues for the possibility of situations in which a collective can be held responsible for its actions, even though none of its members can. He makes use of the so-called Discursive Dilemma to generate this result. This is a situation in which members of a committee have to make a decision about the enactment of a specific policy. In Pettit's example, the committee consists of co-workers who have to decide whether they agree with a pay sacrifice to be used for buying and installing a work floor safety device. Rather than putting that decision directly to a vote, the committee rules specify that the decision will be taken if, and only if, the committee comes to a positive verdict on three issues: whether there is a real danger, whether the safety device is effective, and whether the pay sacrifice is bearable. Furthermore, the rules specify that a vote will be taken on each of the three issues separately and that each issue is decided upon by majority. To simplify matters, assume, as Pettit does, that the committee consists of three members. Table 1 describes their views on the three issues as well on whether the pay sacrifice is indeed justified. Thus each individual is opposed to the pay sacrifice but for different reasons: A does not believe the salary cut is reasonable, B does not think the device is effective, and C does not believe there to be a real danger. However, a vote is taken on each of those issues rather than on the question whether the pay sacrifice should go through. Since for each of the issues a majority exists, the outcome is that the mechanism is installed and the wages cut. However, *none* of the members thinks the pay sacrifice is justified. Can they then be held responsible for the outcome? Consider

Table 1: Employee Safety

	Serious danger?	Effective measure?	Bearable loss?	Pay sacrifice?
A	Yes	Yes	No	No
B	Yes	No	Yes	No
C	No	Yes	Yes	No
Majority	Yes	Yes	Yes	Yes/No

Pettit's view:[28]

> But suppose now that some external parties have a complaint against the group, say, the spouses of the less-well-off workers, who think the pay sacrifice unfair. Whom, if anyone, can they hold responsible and blame for the line taken? Whom can they remonstrate with? Not the individuals in their personal right, since each can point out, the chair included, that he or she was actually against the pay sacrifice and that they were not in a position, as well they may not have been, to see the likely effect of the procedure they followed. The spouses in this example can only blame the corporate group as a whole.

We will set aside Pettit's claim about the status of groups and whether or not they are apposite agents of responsibility ascriptions, because our concern is only whether or not his claim holds with respect to the individual members. To test his claim we again need not introduce the exact details of the game to come to a conclusion. It is easily seen that, by voting as they did, each individual is causally effective for the outcome.[29] We need to focus on the avoidance potential of the various strategies of the individuals. Of crucial importance thereby is whether it is eligible *not* to express one's true beliefs about the issues, for instance when they vote strategically. Now, if strategic voting is not eligible then by Definition 6.1 it is true that *none* of the individuals is to be held responsible – after all, no one has an alternative strategy with a higher avoidance potential. So, in that case, Pettit's claim holds. On the other hand, if it is always eligible to misrepresent one's true beliefs, then *all* individuals are responsible for the outcome that emerged: by voting 'no' on all issues, they could ensure they would never be causally effective for a decision in favour of the wage loss.[30] It is notable that Pettit is silent on the whole eligibility issue; he only explicitly assumes that the members 'vote as they judge', i.e. represent their opinions sincerely; he does make any normative

[28] Copp (2006) and Chapman (2009) have also made use of this quirk of collective decision-making to demonstrate shortfalls in responsibility particular types of committee decisions. For a general review of the Discursive Dilemma, see (List and Pettit 2006).

[29] To avoid any miscomprehension, by the NESS-test there will always be *at least* one member who is a causal factor for *any* configeration of votes.

[30] The only extra assumption that we need to make to get at this conclusion is that any event of two individuals expressing their true beliefs is part of at least some equilibrium.

commitment in this regard. In which case, his claim does not go through, not because he is wrong, but because he has not specified all of the relevant aspects of the situation. Whether a misrepresentation of one's beliefs or convictions is eligible or not may depend on various factors. For instance, if a player herself would see it as a form of immoral behaviour we do not want to demand that she do so. It may also depend on the nature of the issue at hand. In the Discursive Dilemma the issues about which a decision has to be made involve beliefs about the world. In such epistemic contexts we may be more reluctant to say that strategic behaviour is eligible, especially if the members of the committee have been chosen because they are supposed to be experts on the matters at hand.[31] However, in situations in which the disagreements involves values or interests, strategic behaviour may well be eligible – say when we do not vote for our most preferred candidate because we judge that he is less likely to be chosen anyhow.[32]

8. Conclusion

We will not recap the outlines of the paper, but rather conclude with two remarks. The first concerns our methodology and the complex nature of the conditions that we presented. Our necessary and sufficient conditions for ascribing retrospective moral responsibility are informationally rich. It could be argued that this is an argument against our framework. After all, can we really expect that individuals go through the complicated process of establishing all of the ingredients of the game they are in, examine the various causal relations, check the eligibility of their actions, and make the calculations about the avoidance potential of each of their strategies? The question becomes even more pressing when we consider the fact that we have restricted ourselves to settings of complete information. Obviously, this limits the scope of our analysis considerably and a natural next step is to extend the account to situations of incomplete and asymmetric information. However much this may increase the verisimilitude of description this will in all probability make our account even more complicated and the assignment of moral responsibility will be even more difficult.

However, we do not think that the complex nature of the analysis counts against it. As said, we claim to have arrived at the underlying structure of our thinking about an important form of moral responsibility. Hence the title of this paper. The structure is indeed complex, but in discussing two paradigmatic cases in section 7 we have demonstrated that it is *not* necessary to assume that each and every part of it is always invoked when people actually make judgements about who is and who is not responsible for some state of affairs. In both

[31] Hindriks (2009) also challenges Pettit's claim with an argument according to which a misrepresentation of one's beliefs is eligible in an epistemic context. The argument does not refer to the eligibility to act strategically, but to the assumption that expressing a true proposition is always eligible: individuals who have mistaken beliefs about the world may well be demanded to perform a different strategy, regardless of their beliefs, i.e. we can be held to account for our epistemic errors.

[32] Dowding and Hees (2007) have argued that in such cases strategic behaviour is in fact a virtue.

the Tragedy of the Commons and the Discursive Dilemma we did not need all the information about all aspects of the game to arrive at a conclusion. More importantly, we can counter with the argument that people use 'moral heuristics' in the sense of Appiah (2008), i.e. devices through which we arrive at a conclusion that in most of the relevant cases coincides with the one following from a full-fledged analysis of the situation. There is no reason to think our refined definition of moral responsibility (Definition 6.1) will deviate from these heuristics.

Indeed, and now we turn to our second remark, one example of a moral heuristic that approximates our approach – although not the result – is a famous but controversial principle put forward by Singer (1972). In his argument for the claim that we have an obligation to alleviate poverty, Singer stated that '(i)f it is in our power to prevent something bad from happening, without thereby sacrificing anything of comparable moral importance, we ought, morally, do it'. Clearly, this principle is about our moral duties rather than about our (retrospective) responsibility. It is about what we *ought* do – alleviate the plight of people living in destitute circumstances – and not about whether we are accountable for the fact that people live in poverty.

We presented our account as *describing* an important part of much of our practice of assigning praise and blame. However we also think it describes the conditions under which assignments of praise and blame are *justified.* That is, we not only believe that to be responsible for a bad outcome means that one *could* have performed an action that had a higher avoidance potential and which was eligible, but also that one *should* have done so. In this sense, our analysis goes beyond a description of our current practices and presupposes a general principle about what our moral duties are. There is an obvious relation between this general principle underlying our account of responsibility and the principle formulated by Singer. Both invoke a restrictive clause – we say the action should be eligible, whereas Singer states that it should not sacrifice anything of comparable moral worth – and both refer to the necessity of avoiding the bad outcome – we do so in terms of a strategy's avoidance potential, Singer in terms of possible prevention. As we argued, one may be held responsible for some outcome without having had the power to actually *prevent* it. This means that our account of moral responsibility presupposes a view on our moral duties that is even stronger than Singer's view is. Our 'avoidance potential' is not only a descriptive device, but a normative one.

References

Appiah, K. A. (2008). *Experiments in Ethics*. Harvard: Hardvard University Press.

Beebee, H. (2004). Causing and Nothingness. In Collins, J. D., Hall, E. J. and Paul, L. A. (eds), *Causation and Counterfactuals*. MIT Press, 291–308.

Benn, S. I. and Weinstein, W. L. (1971). Being Free to Act, and Being a Free Man. *Mind* 80: 194–211.

Bovens, M. (1998). *The Quest for Responsibility*. Cambridge: Cambridge University Press.

Braham, M. and Hees, M. van (2008). Degrees of Causation. *Erkenntnis* (forthcoming).

Braham, M. and Hees, M. van (2009). Responsibility Gaps and Voids.

Cane, P. (2002). *Responsibility in Law and Morality*. Oxford: Oxford University Press.

Carter, I. (1999). *A Measure of Freedom*. Oxford: Oxford University Press.

Cartwright, W. (2006). Reasons and Selves: Two Accounts of Responsibility in Theory and Practice. *Philosophy, Psychiatry, and Psychology* 13: 143–155.

Chapman, B. (2009). Rational Association and Corporate Responsibility. In Sacconi, L., Blair, M., Freeman, E. and Vercelli, A. (eds), *Corporate Social Responsibility and Corporate Governance: the Contribution of Economic Theory and Related Disciplines*. London: Palgrave.

Copp, D. (2006). On the Agency of Certain Collective Entities: An Argument from "Normative Autonomy". *Midwest Studies in Philosophy* 30: 194–221.

Davidson, D. (1971). Agency. In Binkley, R., Bronaugh, R. and Marras, A. (eds), *Agent, Action, and Reason*. Oxford: Basil Blackwell, 3–25.

Day, J. P. (1977). Threats, Offers, Law, Opinion and Liberty. *American Philosophical Quarterly* 14: 257–272.

Dowding, K. and Hees, M. v. (2007). In Praise of Manipulation. *British Journal of Political Science* 38: 1–15.

Driver, J. (2008). Attributions of Causation and Moral Responsibility. In Sinnott-Armstrong, W. (ed.), *Moral Psychology*. Cambridge, MA: MIT Press, *2*, 423–439.

Feinberg, J. (1968). Collective Responsibility. *Journal of Philosophy* 65: 674–688.

Feinberg, J. (1970). *Doing and Deserving*. Princeton: Princeton University Press.

Fischer, J. M. and Ravizza, M. (1998). *Responsibility and Control*. Cambridge: Cambridge University Press.

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *Journal of Philosophy* 66: 829–839.

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

Goldman, A. I. (1999). Why Citizens Should Vote: A Causal Responsibility Approach. *Social Philosophy and Policy* 16: 201–217.

Halpern, J. Y. and Pearl, J. (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal of Philosophy of Science* 56: 843–887.

Hardin, G. (1968). The Tragedy of the Commons. *Science* 162: 1243–1248.

Hart, H. L. A. and Honoré, A. M. (1959). *Causation in the Law*. Oxford University Press.

Hees, M. van (2008). The Specific Value of Freedom.

Hindriks, F. (2009). Corporate Responsibility and Judgement Aggregation. *Economics and Philosophy* 25: 161–177.

Honoré, A. M. (1995). Necessary and Sufficient Conditions in Tort Law. In Owen, D. (ed.), *Philosophical Foundations of Tort Law*. Oxford University Press, 363–385.

Inwagen, P. van (1978). Ability and Responsibility. *Philosophical Review* 87: 201–224.

Johnson, B. L. (2003). Ethical Obligations in a Tragedy of the Commons. *Environmental Values* 12: 271–287.

Jones, P. and Sugden, R. (1982). Evaluating Choice. *International Review of Law and Economics* 2: 47–65.

Kramer, M. H. (2003). *The Quality of Freedom*. Oxford: Oxford University Press.

Lewis, D. (1973). Causation. *Journal of Philosophy* 70: 556–569, 561: reference to control and power

(Goldman).

Lewis, D. (2004). Causation as Influence. In Collins, J. D., Hall, E. J. and Paul, L. A. (eds), *Causation and Counterfactuals*. MIT Press.

Lewis, H. D. (1948). Collective Responsibility. *Philosophy* 24: 3–18.

List, C. and Pettit, P. (2006). Group Agency and Supervenience. *Southern Journal of Philosophy* 45: 85–105.

Mackie, J. L. (1965). Causes and Conditions. *American Philosophical Quarterly* 2: 245–264.

Mackie, J. L. (1974). *The Cement of the Universe*. Oxford University Press.

May, L. (1992). *Sharing Responsibility*. Chicago: Chicago University Press.

McGrath, S. (2005). Causation by Ommission: A Dilemma. *Philosophical Studies* 123: 125–148.

Miller, D. (2007). *National Responsibility and Global Justice*. Oxford: Oxford University Press.

Morriss, P. (1987). *Power: A Philosophical Analysis*. Manchester University Press.

Nagel, T. (1979). *Mortal Questions*. Cambridge: Cambrisge University Press.

Nozick, R. (1981). *Philosophical Explanations*. Cambridge, MA: Hardvard University Press.

Pettit, P. (2007). Responsibility Incorporated. *Ethics* 117: 171–201.

Sartorio, C. (2004). How to be Responsible for Something Without Causing It. *Philosophical Perspectives* 18: 315–336.

Sartorio, C. (2007). Causation and Responsibility. *Philosophy Compass* 2: 749–765.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Sen, A. K. (1974). Choice, Orderings and Morality. In Körner, S. (ed.), *Practical Reason*. Oxford: Blackwell, 54–67.

Sen, A. K. (1980). Description as Choice. *Oxford Economic Papers* 32: 353–369.

Singer, P. (1972). Famine, Affluence, and Morality. *Philosophy and Public Affairs* 1: 229–243.

Sinnott-Armstrong, W. (2005). It's Not My Fault: Global Warming and Individual Moral Obligations. In Sinnott-Armstrong, W. and Howarth, R. B. (eds), *Perspectives on Climate Change: Science, Economics, Politics, Ethics*. Amsterdam: Elsevier, 285–307.

Stern, N. (2007). *The Economics of Climate Change: The Stern Review*. Cambridge: Cambridge University Press.

Strawson, P. F. (1962). Freedom and Resentment. *Proceedings of the British Academy* 48: 187–211.

Sugden, R. (1998). The Metric of Opportunity. *Economics and Philosophy* 14: 307–337.

Thompson, D. F. (1980). Moral Responsibility of Public Officials: The Problem of Many Hands. *American Political Science Review* 74: 905–916.

Vallentyne, P. (2008). Brute Luck and Responsibility. *Politics, Philosophy and Economics* 7.

Wallace, J. R. (1994). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard.

Watson, G. (2004). *Agency and Answerability*. Oxford: Oxford University Press.

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

Zimmerman, M. J. (1985). Sharing Responsibility. *American Philosophical Quarterly* 22: 115–122.

Zimmerman, M. J. (1988). *An Essay on Moral Responsibility*. New Jersey: Rowman and Littlefield.