An Anatomy of Moral Responsibility

Article *in* Mind · January 2012

DOI: 10.2307/23321778

2 authors:

Matthew Braham
University of Bayreuth
22 PUBLICATIONS **260** CITATIONS
SEE PROFILE

Martin Van Hees
Vrije Universiteit Amsterdam
52 PUBLICATIONS **901** CITATIONS
SEE PROFILE

An Anatomy of Moral Responsibility

Matthew Braham Martin van Hees

February 24, 2010

Abstract

This paper examines the structure of moral responsibility for outcomes both in terms of attributive responsibility and the stronger notion of accountability. We apply game theory to model the anatomy of such responsibility. A central feature of this anatomy is a condition that we term the 'avoidance potential', which gives precision to the idea that moral responsibility implies a reasonable demand that an agent should have acted otherwise. We show how this theory can allocate moral responsibility to individuals in complex collective actions problems, an issue that sometimes goes by the name of 'the problem of many hands'. We also show how this theory allocates moral responsibility in the classic Frankfurt-style example.

1. Introduction

Judgements of moral responsibility for the realization of a state of affairs are intimately connected to the relationship between an agent's behaviour and the state of affairs involved. Roughly speaking, a theory of moral responsibility for outcomes selects a subset of agents deemed to be appropriate candidates for moral appraisal (blame or praise) and/or sanctions (punishment or reward) on the basis of there being a significant connection between their actions and the state of affairs in question. A great many problems fit this general description of moral responsibility, in particular problems that arise when outcomes are related to a complex of agents and decisions.

Consider, first, what is perhaps the most widely discussed example in the literature on moral responsibility, the so-called Frankfurt-example (Frankfurt 1969). It concerns a situation in which an agent performs an action but, had he not decided to perform the action, some intricate mechanism (e.g., a device planted in his brain by another agent) would have made him perform the action anyhow.[1] The example has, of course, been used to call into question the assumption that the existence of alternative possibilities is a necessary condition for ascribing moral responsibility. After all, we feel the agent is responsible for his action even though – given the back-up mechanism – he was not able to do otherwise. Though the example is usually discussed in terms of responsibility for *actions*, it also applies to responsibility for *outcomes*. That is, the example also calls into question whether agents are responsible for the outcomes of their action if the outcomes would have occurred anyhow.

Also relevant for us here is that the outcome depends on what other agents do. That is, the setting is an interactive one – the outcomes of at least some possible decisions depend crucially on the actions of others (here the triggering of the device in the person's brain). Although the Frankfurt-example carries this interdependency between some of the actions of the individuals, in the end it is only the action of one of the two individuals that actually effects the outcome: the back-up mechanism is not triggered. In contrast, there are many real-life situations in which the actions of many different agents *do* in fact contribute to a particular outcome. In these cases ascertaining who is to be held responsible for the resulting outcome is known as the 'problem of many hands' (Thompson 1980). A prominent class of many hands problems are covered by the so-called '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 to the detriment of all the users.[2] In the narrative of the commons, and in collective action problems in general, no individual has direct control over the outcome in the form of actions that are necessary or sufficient

[1] See Widerker (2000) for collection of articles dealing with various forms of these examples.
[2] See Bovens (1998) and Ostrum (1990) for collections of real-world cases.

conditions for such an outcome to occur. In such cases it is not obvious who is to be assigned responsibility for the outcome. To illustrate the high stakes involved, consider the issue of whether moral obligations regarding global warming devolve upon individual people, given that global warming can be traced back to countless day-to-day individual decisions. As many have done, we can describe this situation as a Tragedy of the Commons.[3] In this case the 'commons' is the climate and in one version suggested by Johnson (2003) and Sinnott-Armstrong (2005), the actions involved are, for instance, taking a spin in a gas-guzzling sport utility vehicle (SUV) or not; the outcomes are the many effects, including increased level of carbon dioxide emissions, climate change, and eventual destruction of ecologically sensitive habitats. If we can pin responsibility on Sunday pleasure drivers, at least *prima facie*, an effective policy for the alleviation of global warming should include educative measures that effect changes in the attitudes of the individuals involved. On the other hand, if, as both Johnson and Sinnott-Armstrong believe, such drivers cannot be held responsible for global warning, then any such measures would be unjust in the sense that they place undeserved moral burdens on individuals.

A further class of significant cases concern collective decision making. In a recent and important paper, Pettit (2007) has argued that circumstances in committee voting exist – the so-called 'Discursive Dilemma' – in which none of the constituent members of the committee can be said to be responsible for the outcome of a vote. He christens this 'the problem of no hands'. Pettit considers such circumstances to be particularly onerous because if such responsibility 'voids' exist, then committee members have an incentive to organize their activities in a self-serving way while at the same time being able to avoid any responsibility in the event that a harm occurs (Pettit 2007: 197). As a way around this problem Pettit invokes the concept of collective responsibility. But this carries with it thorny metaphysical and normative implications, yielding the question whether it is indeed impossible to establish individual responsibility in the circumstances he describes.

The objective of this paper is ambitious. We do no less than provide a general analysis

[3]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 modeled as a Tragedy of the Commons.

of the necessary and sufficient conditions for assigning moral responsibility to individuals that can systematically cover the problems of the type that we have just listed. To do so, we will introduce a simple methodological innovation: we take the cases of interactive decision-making as the general case and our starting point. For our analysis of these cases we adopt a framework that is natural to such circumstances, that of game theory. The advantage of doing so is three-fold. Firstly, the results apply to simple and complex decision situations alike because simple decision situations are special cases of the general model. Secondly, it allows us to 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 responsibility in general and cases of joint actions in particular. Thirdly – and we believe most significantly – it abstractly describes the different ingredients of our theory of responsibility. It is for this reason that we say that we present an 'anatomy' of moral responsibility.

It is important to note a limitation of the analysis, a limitation which concerns epistemic considerations in making responsibility judgements. 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 shall assess her responsibility for the outcome differently than if she could not possibly have known, or need not have known, that the outcome would have resulted. We ignore these issues in this paper by restricting ourselves to 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 may *not* have pertains to what the other agents will do; since the consequences of an agent's actions often depend in part on those other actions, the individuals may thus be ignorant about the exact consequences of their actions.

Before beginning we should briefly mention how our analysis is related to the existing literature. There are two main connections. The first one is that a central feature of our anatomy is what we denote as the *avoidance potential*. This is a formalisation of the idea that was recently and independently put forward by McKenna (1997), Wyma (1997), Otsuka (1998), and Hetherington (2003), in the context of the debate about Frankfurt-style

examples. It states that moral responsibility requires that the person in question could escape being blamed or punished for particular outcomes by avoiding being an 'author' of the outcome. Another connection is to the important recent contribution by Vallentyne (2008). Like ourselves, Vallentyne develops a formal framework for analyzing outcome responsibility. There are, however, two crucial differences between his approach and ours. First, the object of Vallentyne's analysis differs from ours. Vallentyne analyzes so-called 'agent responsibility', which is a (possibly non-moral) form of attributive responsibility. We focus on moral attributive responsibility as well as on responsibility as accountability. Accountability entails attributive responsibility, but the converse need not hold. Secondly, whereas we follow Vallentyne in emphasising the importance of there being a causal link between the agent's actions and the outcomes for which he is to be held responsible, our rendition of these causal relations differs substantially from his. One reason is that our informational framework is richer: we assume information is given about the interactive setting in which individuals make their decisions, and then define a non-probabilistic account of causality. Vallentyne abstracts from the background in which the decisions are made and describes causal relations in terms of exogenously given objective probabilities.

As a guide: we start (in Section 2) with demarcating the concepts 'attributive responsibility' and 'responsibility as accountability'. We give an abstract description of two conditions – the Agency Condition and the Causal Relevancy Condition – which together characterizes attributive responsibility, as well as of a third condition – the Avoidance Opportunity Condition – which when added to the other two conditions characterizes responsibility as accountability. Whereas much of the responsibility literature focuses on the kind of agency that is needed for responsibility, we ignore this issue altogether and simply assume that the Agency Condition is always fulfilled. This allows us to focus on where the contribution and innovation of this paper lays: the analysis of the Causal Relevancy Condition and the Avoidance Opportunity Condition.

In Section 3 we introduce and expound upon the basic game theoretic concepts that we will use. Drawing on Braham and van Hees (2009), we introduce and summarize the conception of causal contribution that constitutes the first central component of our

analysis in Section 4. In Section 5 we operationalize our key component of responsibility as accountability: the *avoidance potential*; and in Section 6 we complete our anatomy in full. In Section 7 we apply our formal theory to the analysis of the three puzzles that we listed right at the start. For the Frankfurt case, we show that our analysis supports the conclusion that a person can be morally responsible (both attributively and in terms of accountability) for the realization of a state of affairs, even if there was no course of action available to him that would have led to a different outcome. Although this conclusion is nothing new, the justification is novel: on our account the important consideration is not whether a person could have realized a different outcome but whether he could have avoided making a causal contribution to it. For the Tragedy of the Commons and the Discursive Dilemma we lay bear the conditions that have to be met in order to allocate responsibility to the individual actors. We close with some concluding remarks in Section 8.

2. Moral Responsibility

The concept and cognates of moral responsibility are richly ambiguous and there exists no standard terminology. Our first task is to delineate the conception of moral responsibility for which we will provide a formal framework and to bracket out certain issues.

First, the variety of outcome moral responsibility that is at issue in this analysis is *retrospective*: it is the moral responsibility that an individual bears towards some *realized* outcome or states of affairs.[4] It is about what has happened, such as Jones being killed, the pasture over-grazed, or a particular policy being approved by a committee. Retrospective moral responsibility can be said to come in two forms. The first, and more general, conception is *attributive* responsibility. Roughly speaking this is the notion that an outcome bears the 'authorship', 'hallmark', or 'stamp' of the person in the sense that the behaviour that brought it about is expressive of that person's values and ends. Insofar as we blame or praise people on the basis of their attributive responsibility, we are doing so in view of the

[4]The distinction between outcomes and states of affairs will be clarified in section 3.

person's norms, ends, or character that is reflected in the outcome.[5] We take attributive responsibility to be characterized by the following two conditions.

Agency Condition (AC) The person is an autonomous, intentional, and planning agent who is capable of distinguishing right and wrong and good and bad.

Causal Relevancy Condition (CRC) There should be a causal relation between the action of the agent and the resultant state of affairs.

Although there are different ways of filling in these conditions, on most plausible renditions of attributive responsibility neither of the conditions are sufficient. *AC* is not because we cannot say that an outcome bears the stamp of 'authorship' of a person if the person played no causal role in bringing it about (the outcome may correspond to their goals and norms but that does not mean they 'authored' it) – the child shooting his cap gun did not bring the plane down; and *CRC* is not because we may do many things that effect outcomes but do not bear the mark of 'authorship' – your slipping on the ice that knocks someone over is a cause of their broken arm, but their broken arm does not bear the hallmark of your goals.

A narrower conception of retrospective moral responsibility for outcomes is that of *accountability*. As Watson (2004: 265) has put it, this is a form of responsibility that may permit social censure and sanction in that the agent is accountable for what she has done with reference to a set of shared norms and expectations on the basis of which the censure and sanctions are imposed. This goes further than attributibility because it serves for more than just an appraisal of a person's goals or character; it may, for instance, underpin demands that the person answer for her deeds or claims for remedy (compensation, apology, etc.). To grasp the distinction, consider the bank clerk who hands over the money to a bank robber who threatens to take her life. Suppose that she is attributively responsible, i.e., *AC* and *CRC* are satisfied.[6] Even though attributively responsible she would not,

[5]Here we follow Watson (2004). Note that for Scanlon (1998) attributive responsibility constitutes somewhat more than 'authorship', it is closer to what we term 'accountability' (although not fully so); while Vallentyne (2008) uses the term 'agent responsibility' in place of 'attributive responsibility'.

[6]Of course, whether this is indeed so depends on how the conditions *AC* and *CRC* are further filled in.

usually, be considered as accountable for it. Assuming the threat was credible, we will not ask her to answer for her deeds let alone impose censure and sanctions for not having complied with the demands.

The question is, then, what additional condition narrows moral responsibility down to accountability? A common line of thought is that some form of *control* other than being a causal factor is required. In Watson's (2004: 280) view, because blaming responses potentially affect the interests of those blamed, moral accountability is closely related to the issue of *avoidability*. We postulate:

Avoidance Opportunity Condition (AOC) The agent should have had a reasonable opportunity to have done otherwise.

The question, now, is how this avoidability is to be understood. Note that we do not say that the agent should have had a reasonable opportunity to realize an alternative outcome; the condition is formulated in terms of alternative actions rather than outcomes. We often hold individuals accountable for some outcome even though they could not, on their own, prevent it from occurring. The rationale for this can be highlighted by two very simple examples involving causal overdetermination.

Consider, first, the canonical example of Two Assassins. Two assassins, in place as snipers, shoot and kill Victim, with each of the bullets fatally piercing Victim's heart at exactly the same moment. Clearly neither assassin could have changed the outcome (Victim dies) by not shooting. Yet we do not want to say that they are, as a result, unaccountable for Victim's death. For the second example, consider a committee that makes its decision on the basis of the majority rule. If the members voted unanimously in favour of some proposal, then clearly none of the individuals could have changed the outcome by voting differently. As in Two Assassins, we do not want to say that the committee members

On a weak interpretation of AC, for instance, one could maintain that despite the duress the clerk under, the outcome bears the hallmark of at least one of her goals – maintaining her bodily integrity. Stronger interpretations of AC may also yield the conclusion that the clerk is attributively responsible but the story may then have to be changed somewhat: say the clerk knew the robber and wanted him to have the money irrespective of the threat.

are unaccountable for the decision to pass the proposal. Hence, being able to unilaterally control outcomes is too strong a condition for accountability.

Our account of the required avoidance opportunity is motivated by the views developed recently and independently by McKenna (1997), Wyma (1997), Otsuka (1998), and Hetherington (2003). Ignoring the details of their accounts and stating it in the terminology we use here, their views can be summarized by the principle that *one can only be accountable if one had reasonable opportunity not to be attributively responsible*. The nub of the idea is that a person is an apposite target of censure and sanction for some state of affairs only if, to paraphrase McKenna (1997), it was within her power to reasonably choose *not* to be an 'author' (or 'co-author' in the presence of multiple causal conditions) of that state of affairs. In the bank robber example, we do not submit the bank clerk to, say, censure and sanction for co-authoring the state of affairs in which the bank robber walks out with the loot because she did not have a reasonable opportunity not to be such a co-author.

The conditions for attributive responsibility that we lay out here allow us to give further precision to this principle. Clearly, there are two types of circumstances in which a person is *not* attributively responsible: those in which AC is not met and those in which CRC is not met. An important caveat here is that we will assume – but without any justification – that under the full information setting we presuppose (see the introduction) an autonomous decision to act non-autonomously never undermines one's responsibility for the consequences of the non-autonomously performed actions. Simply put, if we know the consequences of excessive drinking, then, if we deliberately become drunk, our drunkenness does not undermine our accountability for those consequences.

The required avoidability therefore comes down to having the possibility to avoid meeting CRC: it consists of the possibility to break the causal link between the agent and the outcome. To illustrate, in Two Assassins, by not shooting, an assassin would not have prevented the outcome, but would have ensured that he would not be causally effective for it (we will return to this example in Section 4). Similarly, for the committee example, by voting against the proposal an individual committee member would not have prevented the outcome, but would have ensured that he would not be causally effective for it being

approved.

Before proceeding, we recognize that all three of these conditions carry with them their own special problems. For the purposes of this analysis we will simply assume that AC is always satisfied, which means the behaviours we analyse do indeed have the appropriate quality; that is, the individuals in our analysis are assumed to have a capacity to reflect on their beliefs and desires and exercise some form of control over what they do, or refrain from doing: they are 'reason-responsive' and 'planning agents' in the required sense. Ignoring how this capacity has to be filled in allows us to concentrate on fleshing out and giving formal and abstract structure to the second and third condition. It is here where the innovations lie.

3. The Formal Framework

The elementary concept that we need throughout is that of a *game*. Very generally, a game specifies the 'rules of the game', which define precisely who can do what, when, and to what effect, as well as the individuals' preferences regarding the various outcomes. We use the information provided by a game as the basis of responsibility assignments. Formally, 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 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.

The elements of the set X are taken to be social states, viz., detailed descriptions of the world. While the elements of X are called *outcomes*, the (non-empty) 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 extensionally. Hence, we have $x \in A$ and $A \cap B = \emptyset$.[7] The distinction between states of affairs and outcomes is important. A person may be responsible for an outcome without being so for all of its aspects (I may be responsible for Bob being elected but I need not be so for the fact that a man is elected, as there may have been no female candidates on the ballot); conversely, one may be responsible for a state of affairs without being so for the outcome (I may be responsible for the election of a male candidate – perhaps because I vetoed all female candidates – without being so for Bob's election).

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 a gun, aiming, curling a finger around the trigger, etc. In general terms it is a plan of action or a protocol of when to do what. A game also treats 'omissions' as strategies. A strategy of 'not shooting', for instance, is any set of actions that does not contain certain actions that 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 a particular outcome to each combination of strategies. The function is completely abstract: it may be determined by empirically ascertainable 'laws of nature', such as the strength required, say, to lift a fallen tree that has trapped a person; or it may be a matter of social law and convention, such as a decision rule that determines who has the authority to do so certain things. In any case it can be a statement of certain regularities, generated either 'naturally' or 'conventionally'.

In itself, a specification of the players, their strategies, the outcomes, and the outcome function 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*,

[7] To avoid cumbersome notation, we shall 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.

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 give a consistent ordering of their wants, desires, needs, and interests. A solution concept specifies how rational individuals will play the game: it consists of a subset of plays. For our purposes it is important not only to know what the agents will do given the game as described and the assumption of individual rationality in place, but also what would happen if a player were to adopt a strategy that is not rationalizable. For this reason, we assume that for each player i and each strategy s_i available to her, we can assign a probability distribution over the plays in which i performs s_i. We assume these probabilities to describe justified expectations about the possible actions of the others. We do not further discuss the precise nature of these probability distributions, or how they are derived from a solution concept.[8]

Before we proceed, it is important to point out that the specification of the relevant game is not an innocuous choice. To see this, consider Beebee's (2004) 'Queen of England Problem'. Suppose Bob promised to water Joan's plant but for some reason 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 Bob and Joan are the only players. Now suppose we derive the judgement that Bob 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 led to the judgement that Bob's omission is a causal factor may also apply to the Queen's omission.[9] However, the Queen is not assumed to be part of the game and thus the way 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 broadly adopt, the

[8]Note that the probability distributions are defined for any s_i, hence also for those that have a zero probability of being adopted by i. See Brandenburger (2007) for a detailed presentation and discussion of the use of such probability functions in a game theoretic framework.

[9]On the account of causation that we present in the next section this is indeed the case. After all, had Bob 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).

ingredients of the analysis are what he calls the 'causal field'.[10] This notion presupposes knowledge of 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', Bob'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'; the game contains all of the normatively relevant information.

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 eventual outcome. 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 responsible after all – perhaps because if she had not agreed to join the committee the vote may not have been taken at all. In the foregoing analysis we will indeed focus only on a *given game* at hand, which means we will not handle moral responsibility for past choices.

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 are often interested in assigning responsibility in well-defined and specific contexts, such as when we want to know who is responsible for a specific committee decision, insofar as that responsibility can be traced back to the decision-making process *within* the committee. In these cases, the research question at least in part determines the moral field. Secondly, the relevance of the moral field points out that assignments of moral responsibility are partly determined by

[10]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.

our prior moral expectations: there is, we assume, no normatively neutral way of arriving at such judgements. This means that the choice of the game should be carefully justified; it does not mean that such justification is impossible.[11]

4. Making a Causal Contribution

In order to ascribe moral responsibility for some state of affairs to an individual we must precisely parse her causal connection to that state of affairs. This is the condition *CRC* stated in Section 2. Two remarks are necessary before we begin. First, although we use the term 'causation' or ' causal efficacy', we are aware that on occasion it would be 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 'law of nature'. The cases that we 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 and intensity 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'.[12] To adopt this other terminology would, however, burden the discussion without adding anything. In fact, we need not do so since we are primarily concerned with causation for *particular* or *singular* events (or 'causes in fact' in legal terminology).[13] The conception of a cause that we adopt, but do 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

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

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

[13]For the reader unfamiliar 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.

what would have been.'

We assume, as is general in legal theory (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' with 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.[14]

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 (NESS). Or, somewhat more precisely, C is part of a set of conditions that are together sufficient for E and is necessary for that set of conditions to be sufficient for E. In the case in which all elements of the set are necessary for the sufficiency of the set, we call such a set a *minimal sufficient condition* for E.[15]

For our purposes it is helpful to note here that the NESS test readily accounts for cases of causal overdetermination. The reason for this is that an event is attributed causal status

[14]The NESS test was 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) provides 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 need the apparatus here, however. As a point of contrast, in his formal framework Vallentyne (2008: 63) makes use of an entirely different conception of a cause. He assumes that 'the relevant causal connection is that the choice increases the objective chance that the outcome will occur, where objective chances are understood as objective probabilities in the sense of single case propensities.' We discuss reasons for not adopting this model in Braham and van Hees (2009).

[15]Sometimes the NESS test is restricted by imposing the additional requirement that the sufficient sets must always be minimally sufficient (see Mackie (1965, 1974), Wright (1988), and Halpern and Pearl (2005)). We do not use this strengthened version, however, because it suffers from a number of paradoxical problems, which are discussed in Braham and van Hees (2009).

even if, due to the presence of other actually or hypothetically sufficient sets, it was not necessary for the result in the relevant circumstances. To see how this works, recall the case of Two Assassins (Section 2). There are two sets of minimally sufficient conditions Victim's death that were present on the occasion, viz., each assassins shot. Thus each assassin contributed a causal factor for Victim's death. Consider another case: Three Walkers. 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. Here there are three possible sets of actions that are minimally sufficient for the rescue and each rescuer has performed an action that belongs to at least one (in fact two) of these. Thus, each of the rescuers' actions can be attributed causal status.[16]

To formulate this notion of causality in game-theoretic terms we 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); we shall write s_{-i} or s_{-T} instead of $s_{N-\{i\}}$ or s_{N-T}, respectively. 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_{-T}) to denote the play of G that consists of the combination of the (mutually exclusive) events s_T and s_{-T}. We let $\pi(s_T)$ denote the set of outcomes that can result from the event s_T: $\pi(s_T) = \{\pi(s_T, s_{-T}) \mid s_{-T} \in \times_{i \notin T} S_i\}$.[17] 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. The game-theoretic rendition of the NESS test becomes:

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

[16] More examples are discussed in Braham and van Hees (2009).

[17] 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 $\pi(s_T) = \{x\}$.

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'_{-T} for the players outside of T such that $\pi(s'_i, s_{T-\{i\}}, s'_{-T}) \notin A$. Thus, according to our account, being 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 possibility alone is inappropriate for constituting moral responsibility.

Before proceeding we need to acknowledge a limitation that arises from our focus on games in normal form. In such games, the NESS test has problems dealing with some types of strategies, viz., those comprising conditional actions. Suppose Bob adopts the strategy of painting his garage door if, and only if, his son does not do so. If we were to model the situation as a normal form game, Bob's strategy would be said to be a NESS condition for having the door painted: any outcome resulting from it would lead, by assumption, to the door being painted. Bob would thus be said to have made a causal contribution to the outcome even when it turns out that he was reading the newspaper as his son painted the door. For cases in which such conditional strategies exist, we should apply the NESS test to the underlying games in extensive form, rather than to the game in normal form. We do not go into this issue here and except for our discussion in Section 7.3 we assume strategies comprise non-conditional actions only.

5. Having a Reasonable Opportunity to Do Otherwise

Like the Causal Relevancy Condition (CRC), the Avoidance Opportunity Condition (AOC) requires filling out. As we said, we take it as given that for a person to be a legitimate target of moral censure and sanction they must have had a reasonable opportunity to have avoided (co-)authorship of the outcome, and this opportunity must be reasonable by some

criteria.

Similar to our handling of *CRC*, we proceed abstractly. We take a *mere* opportunity to be a strategy *simplicter*, and *reasonable* opportunity to be a strategy of a particular kind, which, in our our view has two properties. The first property refers to the possible consequences of adopting that strategy. Since those consequences are at least partly affected by what the others do, we should take account of these strategic interdependencies. For this purpose we introduce the notion of a strategy's *avoidance potential*, which is the potential for avoiding authorship. The second property is that the strategy is *eligible* by some standard.

To define this first property, the avoidance potential, we assume the existence of a family of subjective probability distributions describing the justified beliefs (see Section 3). Using our formal apparatus, we say that for any game G and any i, let $p(s_{-i} \mid s_i)$ denote the probability of the others playing s_{-i} if i were to adopt s_i. Furthermore, let $h(A, s_i)$ denote the set of strategy combinations s_{-i} such that i is not causally effective for the realization of A in (s_i, s_{-i}). We now have:

Definition 5.1 The avoidance potential, $\rho_i(s_i, A)$, of a strategy $s_i \in S_i$ for a state of affairs A, equals

$$\sum_{s_{-i} \in h(A, s_i)} p(s_{-i} \mid s_i).$$

The avoidance potential should not be confused with the idea of control over outcomes (such as in van Inwagen's (1978) 'Principle of Possible Prevention'). What is crucial to the strategy's avoidance potential is that we do not examine whether it may yield *a state of affairs different from* A but, as we have said earlier, whether it *fails to be a causal factor* (by the NESS test) in 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 condition of 'control' would establish a causal link to an alternative outcome.

To get to grips with the avoidance potential, consider once more the case of Two

Assassins. It now is easy to see why both Assassins are indeed accountable for Victim's death. Consider Assassin 1. We know from Section 4 that just like Assassin 2, he is causally effective for it by way of the NESS test. Hence, assuming that AC is satisfied, he is attributively responsible. Even though he could not have prevented Victim's death, he is accountable. The crucial point is that he could have avoided being a causal factor for the realization of Victim's death: if he had not shot, he would not have been causally effective. Of course the same argument applies *in toto* to Assassin 2.[18]

As such, the avoidance potential is still too raw and this brings us to the second property. A strategy strategy may increase the probability of avoiding a particular outcome but it may not be 'eligible' in the sense that there are reasons why it should not be performed. To fix ideas about this requirement recall the example of the bank clerk (Section 2). Even though the avoidance potential for not handing over the money (suppose she could press an emergency button and lock the doors) is lower than handing over the money, we do not want to hold the bank clerk accountable for the bank robbery – given that her life was at risk, it would not have been appropriate to demand that she refuse to comply with the demands of the robber. Similarly, and less grimly, we want to be able to distinguish between criticising someone causing pollution by taking Sunday pleasure drives – 'don't be a spoilsport, I'm having such fun' – and someone causing the same amount of pollution by driving someone to a hospital – 'don't be ridiculous, my wife was in labour'.

The avoidance potential thus needs to be refined in a specific way. The alternative strategies that make up the potential should be 'eligible' or 'acceptable' according to some standard. We use the term *eligibility*, which we take from the freedom literature. This 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

[18] As a point of comparison, Vallentyne views neither of the two assassins (attributively) responsible for Victim's death since the outcome occurs no matter what the other does. Vallentyne, however, does ascribe attributive responsibility to each assassin for the *act* of shooting.

freedom consists of the possibility that an agent performs 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, in Benn and Weinstein's (1971: 195) opinion, being 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)'.

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 affecting our responsibility.[19] The underlying intuition is that in the same way as an ineligible opportunity should not be considered to have a substantial effect on our freedom, the availability of such a strategy should not be considered to affect our responsibility. In other words, a person is to be excused for the realization of some bad outcome, or does not deserve praise for a good one, if its avoidance was only feasible by the adoption of an ineligible strategy.

Obviously the question is what constitutes an eligible strategy. For now we shall simply presuppose the existence of some theory about what an eligible strategy means and give the conception formal content as follows:[20] for each individual i, eligibility is represented by a mapping $\mathcal{E}_i$ that assigns to each contingency s_{-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. If the $N-i$ players, for instance, perform actions that would

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

[20]Eligibility could be welfaristic in the sense that strategies should be welfare maximizing by some measure; or they could be deontic in the sense that do not violate basic obligations. For our purposes we need not commit ourselves to any particular position. It is important to note that different positions may yield different responsibility judgements. Consider the following possible objection raised by a referee. A can avoid doing a small amount of damage to B's car, which has stalled on the road, only by doing a significantly larger amount of damage to her's by swerving into a ditch. Even though A had no reasonable opportunity to do otherwise, some may still believe she is accountable for the damage done to Bs car. On our view, rather than forming a counterexample, the example hints at the possibility of conflicting views on the proper normative basis of our eligibility judgements. On a purely consequentialist account of eligibility, A's driving into B's car may be acceptable and therefore she is not accountable for the damage she inflicted. On the other hand, under an extreme deontological view A could be morally accountable because she has a moral obligation never to harm others, an obligation which holds also when the harm is avoids greater harm to one's own property.

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 here in fleshing out a normative theory (the substance of 'eligible alternatives'), but only in investigating structural features – the anatomy – of a particular form of moral responsibility in games, we refrain from any further exploration of the nature of these eligibility functions.[21]

6. Formal 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 as accountability. It consists of three formal conditions that together are necessary and sufficient for selecting the set of individuals who are legitimate objects of moral censure and sanction. Call a 'responsibility game' a triple $(G, \Gamma, \mathcal{E})$, where G is a game, Γ the family of probability functions (one for each strategy of each individual) associated with a particular solution concept and reflecting the justified beliefs of the agents, and $\mathcal{E}$ a set of individual eligibility functions (one for each individual).

Definition 6.1 In a play (s_i, s_{-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. For all $i \in N$, i is an agent in the relevant sense (i.e., AC is assumed to be satisfied);
2. s_i was a causal factor for the realization of A;
3. there is some s_i' in $\mathcal{E}_i(s_{-i})$ 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 interactive decision-making.

The definition is a 'result' in itself because it provides a clear set of variables and a

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

relation between these variables for discussing moral responsibility as accountability. That is, the definition informs us about the data and base-relations that need to be accounted 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 account of the full range of variables that we have introduced in our definition. So, for instance, none of the examples that are discussed in either Fischer and Ravizza (1998) or Sartorio (2007) mention preferences and solution concepts; and while these are mentioned in Johnson (2003) and Sen (1974), their discussions are basically restricted to eligibility considerations. 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 the belief structures (the probability distributions) that are an integral part of the avoidance potential into account.[22] In our approach all these variables are considered (e.g., preferences enter the equation via the avoidance potential). It is also worth noting that the definition is silent about socially optimal states of affairs. No burden is placed on a player to necessarily choose the socially optimal outcome (by some standard) and then hold the player responsible for its non-occurrence 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. We have not taken a stand on whether whether moral responsibility can be quantified or distributed in some particular way, and nor will we do so here. 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.[23] 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

[22]Goldman's solution is in fact a special case of Definition 6.1 in which he implicitly assigns equi-probability to each possible strategy play.

[23]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 that such statements are meaningful. See Zimmerman (1985) for an argument against the notion that moral responsibility can be expressed in terms of relative shares of a fixed sum.

quantifications.[24]

Finally, we note that the definition has both descriptive and normative content. On the one hand we claim that *any* definition of moral responsibility that connects outcomes to individuals will, one way or another, take this form. That is, any ascription of this form of 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 moral responsibility *ought* to take this form, i.e., our constructive exercise demonstrates the proper way to make such ascriptions.

7. Applications

With the machinery in place, our task now is to evaluate the three examples that we used to motivate our analysis in Section 1. We focus first on the general collective action cases before turning to the more difficult and subtle Frankfurt example. To keep the analysis tractable we will avoid, as far as possible, the underlying formalisms that we outlined in earlier sections, although this underlying formalism is what generates our answers.

7.1 *The Tragedy of the Commons*

As we mentioned at the outset, the 'Tragedy of the Commons' (Hardin 1968) is a model that captures a whole host of important collective action situations (global warming, depletion of fish stocks, production of public goods, volunteering). To analyze whom is to be held responsible for the depletion of the commons we could decide to present the game in detail, state the solution concept and corresponding probability functions, 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

[24] In Braham and van Hees (2009) we demonstrate how such quantifications are possible for purely causal contributions and therefore for conceptions of responsibility that are defined purely in these terms. If AC is satisfied then the measure we propose could be interpreted as a measure of attributive responsibility. See also Felsenthal and Machover (2009).

restrict ourselves to three general characteristics that we assume all such situations to have in common. (i) The bad outcome could have been prevented, i.e., the game in question has at least one other outcome. (ii) Before the game was played, each agent assigned at least some positive probability to the strategy combination the others actually did play. (iii) Regardless of the strategies of the others, each individual had at least one eligible strategy by which he would not be effective for the outcome.

The alternative strategy by which an individual would not have been effective is usually taken to be some sort of restriction on the part of the player, such as a restriction of the number of livestock a farmer brings to the commons, a reduction of one's 'carbon footprint' activities, etc. Note that 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 livestock to the commons means that a farmer and his family will starve to death for lack of income or food, then it is not reasonable to demand that he do so. We only assume that *some* strategy of restricting oneself is eligible.

Given these characteristics, it is not only easy to see that *some* individuals are responsible for the outcome but also that *all* individuals who causally contributed to the outcome are so. That at least some of the individuals are causally effective for the outcome follows from the first characteristic. That is, it entails that some subset of the strategies actually played forms a minimal sufficient condition for the outcome. Now take any of the players whose strategy is a part of such a minimally sufficient condition. The second characteristic states that the player had (a justified true) belief that he might be causally effective for the outcome. Since, by the third characteristic, he could have ensured that he would not make a causal contribution, and since doing so would have been eligible, it follows that he is morally responsible (as accountability).

We infer that an excuse of the form 'given that I could reasonably – and correctly – have expected each of the other agents to play their polluting strategy I cannot be blamed for doing so too because my contribution to the depletion of the resource would not have any noticeable effect' will not pass. In other words, a failure of collective action does not have a

moral correlate in the form of a failure of individual responsibility, as Hardin (1988: 156–7) appears to suggest. And thus in contrast to, say, Johnson (2003) and Sinnott-Armstrong (2005), our framework will assign responsibility for global warming and its manifold effects from frivolous polluting activities.

7.2 *The Discursive Dilemma*

Our next application concerns the recent argument put forward by Pettit (2007) in which he claims that situations exist in which a collective can be held responsible for its actions, even though none of its members can.[25] 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 by majority vote. 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 as whether the pay sacrifice is indeed justified.

Each individual is thus 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 is a real danger. However, a vote is taken on each of those issues rather than on the question of whether the pay sacrifice should go through. Since a majority exists for each of the issues, the outcome is that the mechanism is installed and the wages

[25]We thus take what we call a 'responsibility void' to be a situation in which none of the individuals bear any responsibility for the outcome. In fairness to Pettit, it should be noted that he leaves open the possibility that responsibility voids need not consist of such a complete shortfall in individual responsibility but may already exist if each of the committee members bear 'only a very diminished level of personal responsibility for the outcome' (p. 198). Since we do not derive judgements about degrees of responsibility, our analysis of the Discursive Dilemma does not say anything about such partial responsibility voids.

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

cut. However, *none* of the members thinks the pay sacrifice is justified. Can they then be held responsible for the outcome? Consider Pettit's position on the matter:[26]

> 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 a responsibility ascription, 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 easy to see that, by voting as they did, each individual made a causal contribution to the outcome. We thus need to focus only on the avoidance potential of the various strategies of the individuals. Of crucial importance here is whether it is eligible *not* to express one's true beliefs about the issues in the form of votes, as 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

[26]Copp (2006) and Chapman (2009) have also made use of this quirk of collective decision making to demonstrate shortfalls in responsibility for particular types of committee decisions. See List and Puppe (2009) for a general overview of the Discursive Dilemma.

responsible – after all, no individual then has an alternative strategy which is eligible, and thus also does not have an eligible strategy with a higher avoidance potential. So, in that case, Pettit's claim holds.

On the other hand, if all of the members assigned at least a positive probability to the others voting sincerely, the eligibility of the misrepresentation of one's true beliefs would imply that *all* individuals are responsible for the outcome that emerged: by voting 'no' on all issues they could have ensured they would never be the causal factor for a decision in favour of the loss of wages. So, in that case, there is no responsibility void. Pettit does not discuss the eligibility issue. He only goes as far as to say that he assumes that the members 'vote as they judge', i.e., they represent their opinions sincerely, but without making any normative commitment in this regard. In which case Pettit's claim does not go through, not because he is wrong but because he has not specified all of the relevant aspects of the situation. [27]

Whether or not a misrepresentation of one's beliefs or convictions is eligible may depend on a variety of factors. For instance, if a player would personally consider it a form of immoral behaviour we may not want to demand that she do so. It may also depend on the nature of the issue under consideration. 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 factual matters being considered. However, in situations in which the disagreements involve 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.[28]

[27] The same conclusion holds for Copp's (2006) analysis of moral responsibility in the Discursive Dilemma. We cannot comment on Chapman's (2009) because he is concerned with legal rather than moral responsibility so that Definition 6.1 may no longer be appropriate. For an alternative account of moral responsibility in the Discursive Dilemma, see Hindriks (2009).

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

7.3 *Frankfurt-style Examples*

Frankfurt's (1969) example is of an entirely different quality to the two cases we have just examined. It is far more abstract as it is about testing the very idea that the existence of alternative possibilities is a prerequisite for the assignment of moral responsibility. To recap, in Frankfurt's example, an agent called Jones performs some action a leading to an outcome x, say he kills Smith; and he does so in an entirely premeditated way and for his own selfish reasons without any prompting from a third party. In our terms, Smith's death bore the 'stamp' or 'authorship' of Jones. However, unbeknownst to Jones was the presence of Black who also wanted Smith dead and had a way of manipulating Jones's decision such that if Jones were to decide *not* to perform the action leading to Smith's death Black could intervene such that Jones would perform the action after all. But given the circumstances, Black did not have to intervene. Intuition seems to point in the direction that Jones is morally accountable for Smith's death despite the fact that he did not seem to have any opportunity to realize a different outcome. How do we justify this accountability?

Applying our framework to the scenario seems problematic for two reasons. First, we have restricted ourselves to settings of complete information, whereas the usual reading of the Frankfurt example is of one incomplete information: Jones does not know that Black can interfere with his decision making. However, as a starting point, there is nothing in Frankfurt's discussion of the case that gives us reason to not to suppose he has this information, i.e., that the game is one of complete information. Furthermore, as we will point out, for this example, the informational setting will turn out to be irrelevant. A second possible complication is the one we mentioned in Section 3, which is that our game-theoretic account of causation only works for strategies that are non-conditional – if a strategy has a conditional nature, the application of the NESS test may yield counterintuitive results. Black's option of interfering if Jones were to decide not to perform kill Smith is, however, a conditional strategy. Fortunately, we can ignore this complication as well because it only relates to the application of the NESS test to Black's actions; it does not affect the analysis of Jones's causal contribution. Yet it is Jones's situation that we want to analyze in terms

of the avoidance condition.

To model the situation game-theoretically we need to describe the strategies of Black and Jones. Black's strategies are clear. We can take him either to adopt the conditional strategy of interfering with Jones (denoted by t_1) or the unconditional one of not doing anything (t_2). With respect to Jones's strategy set, things are more complicated. One of his strategies (denoted by s_1) is the decision to perform a – that much is clear. However, since Black can intervene *if* Jones were *not* to decide to do a, we can describe his second strategy (s_2) as the action of *not deciding* (loosely speaking *intending* or *trying*) to do a. This yields Figure 7.1 as a partial description of the game, with y being the outcome in which Smith lives.[29]

		Black	
		t_1	t_2
Jones	s_1	x	x
	s_2	x	y

Figure 7.1

To assess Jones's responsibility in (s_1, t_1), which is the way the game is played in the narrative, we assume that adopting s_2 would have been an eligible strategy for Jones. By assumption, Jones's actual strategy s_1 (doing a) is a sufficient condition for the realization of x, whereas s_2 (deciding or intending not to do a) is not. Hence, application of the NESS test shows that Jones makes a causal contribution to x. Furthermore, in any of the plays in which he decides to play s_2 (decides *not* to do a), he does not make a causal contribution to the realization of x – either because x is not realized (Black does not interfere) or because x is realized (Black intervenes) but Jones's strategy fails to pass the NESS test. The upshot is that the strategy s_2 of deciding not to perform a has a higher avoidance potential for x than s_1, the strategy that he played, and consequently Jones is accountable for Smith's death. Note that from the game form it is easy to see that this result holds even if the

[29]The matrix only defines part of the game since it does not give any information about the agents' preferences.

complete-information assumption is not satisfied (Jones is unaware of Black's presence). In that case the avoidance potential of s_2 also is higher than s_1, and Jones knows this. The only difference is that Jones may have a wrong belief about *why* he will not be causally effective for x when he plays s_2 - but we can safely assume that the wrongness of that belief does not undermine his accountability if he adopts s_1.

Our analysis supports Frankfurt's conclusion that a person can be morally responsible for the realization of a state of affairs, even if he did not have a course of action that would have led to a different outcome. On our account, however, Jones is to be held responsible for the realization of x because he performed a as an agent, because he made a causal contribution *and* because he had sufficient avoidance potential. Frankfurt's argument is different. He claims that the unactualized scenario in which Black intervenes does not play a role in the responsibility assessment; it can be subtracted from the assessment of Jones's moral responsibility for x. What counts is whether or not Jones 'really wanted' to perform a to obtain x.[30]

The argument that we use here, and more specifically the kind of alternative possibility that we attribute to Jones, has been developed and defended variously by McKenna (1997), Wyma (1997), Otsuka (1998), and Hetherington (2003). The argument has been pointedly criticized by Fischer (1999: 117ff) who claims that it pays no heed to the intuitive requirement that any alternative opportunity must be sufficiently *robust*:

> Even if there exists the possibility that the agent not be the author of his action (or avoid moral responsibility), it does not follow that the agent has the ability (in the relevant sense) to avoid authorship (or responsibility) (p. 122).

What Fischer has in mind here is his criticism of the 'flickers of freedom' response to Frankfurt's example. The flicker theorist argues that alternative possibilities do exist in the Frankfurt-scenario because although Jones cannot choose not to perform a, 'his brain can still exhibit a different neurological pattern N^* (from the one he actually exhibits, N)' (p. 110). For the flicker theorist moral responsibility can be traced back to some suitably

[30]This thesis is developed further in Frankfurt (1971).

placed flicker of freedom indicating an intention that can be read by a Black-type agent as the basis of his intervention. Fischer is of the opinion that such flickers are simply 'too thin a reed to rest moral responsibility' (p. 110).

We are not moved by this criticism. After all, the alternative possibility consists of a *decision* that leads to the sign (neurological or otherwise); it is not the display of the sign itself (c.f., McKenna (1997: 74–5) and Wyma (1997: 64)). Now, Fischer also argues that the defense will fail on the grounds that

> ...in the alternative scenario in a Frankfurt-type case, the agent does not choose to escape responsibility or voluntarily choose anything which implies his escaping responsibility. To bring this point out a bit more clearly, note that in the alternative scenario in a Frankfurt-type case the agent does not deliberate about whether or not to embrace moral responsibility. Issues about whether or not to be morally responsible play no explicit role in his deliberations (p. 121).

But this, too, fails to convince. Fischer is making two claims here, one of which is untrue, the other irrelevant. The first claim is that in the alternative sequence in which Black intervenes Jones does not voluntarily choose to do anything that implies his escaping responsibility. The problem here is that Fischer has the wrong object in his sights: even though performance of a (killing Smith) is indeed involuntary, the *decision* or *intention* not to to shoot *is* a voluntary one. And it is this event that we take to be expressing the ability in the 'relevant' and 'robust' sense on which we should, and do, base judgements of moral responsibility

The other point that Fischer raises concerns the fact that in the alternative sequence considerations of moral responsibility play no role in Jones's deliberation. We agree but do not believe it to be relevant. Suppose Jones wanted Smith dead but decided not to kill him because it would be too much of a hassle to get out of his chair and get the gun from his attic. Black now intervenes: Jones gets his gun and kills Smith. In this case we do not hold Jones accountable – even though considerations of evading moral responsibility

played no role in his deliberation. What is crucial is the absence of the decision to try to kill Smith, not the reason why it is absent.

8. Conclusion

We end with two final remarks. First, our necessary and sufficient conditions for ascribing moral responsibility are informationally demanding. It could be argued that this is an argument against our framework. After all, can we really expect individuals to 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 do 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, it will in all probability make our account even more complicated and the assignment of responsibility more difficult.

We do not believe that the complex nature of the analysis is a valid criticism. As stated, 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 the paradigmatic cases of Section 7 we have demonstrated that it is *not* necessary to invoke each and every part of it to arrive at make judgements who is, and who is not, responsible for some state of affairs.

Second, our formal framework provides a method to compare and assess different conceptions of outcome responsibility in the form of a specified set of variables and their relations – the set of players, their strategies, the outcome function, the causal relations, the avoidance potential, and eligibility. As we have intimated, attributive conceptions are based on the agency and causal conditions, and different attributive conceptions will differ according to the formal content of these conditions. A prime example is the difference that

we have cited between Vallentyne (2008) and ourselves on this score: he fills out the causal condition in terms of (objective) probabilities; we fill it out with with the NESS test. In a similar vein of thought different conceptions of the Principle of Alternative Possibilities can be compared in this framework. One view, which following Fischer and Ravizza (1998) may be called the regulative control version, would require that an agent has strategy that can force an alternative outcome. We agree with Fischer and Ravizza that this condition is much too strong. Instead, we have argued for the view that the relevant control is given by a strategy that breaks the causal link with the actual outcome. In doing so, we have been exclusively concerned with the causal and avoidance conditions and simply took on-board in a very general way the agency condition. A complete anatomy of moral responsibility would require that we give formal content to this condition as well.

References

Beebee, H. (2004). Causing and Nothingness. In Collins, J. D., Hall, E. J. and Paul, L. A. (eds), *Causation and Counterfactuals*. Cambridge, MA: 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 (2009). Degrees of Causation. *Erkenntnis* (Online First).

Brandenburger, A. (2007). The Power of Paradox: Some Recent Developments in Interactive Epistemology. *International Journal of Game Theory* 35: 465–492.

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

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

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.

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

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

Felsenthal, D. S. and Machover, M. (2009). A Note on Measuring Voter's Responsibility. *Homo Oeconomicus* 26 (forthcoming).

Fischer, J. M. (1999). Recent Work on Moral Responsibility. *Ethics* 10: 93–139.

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.

Hardin, R. (1988). *Morality within the Limits of Reason*. Chicago: Chicago University Press.

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

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

Hetherington, S. (2003). Alternate Possibilities and Avoidable Moral Responsibility. *American Philosophical Quarterly* 40: 229–239.

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.

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

List, C. and Puppe, C. (2009). Judgement Aggregation: A Survey. In Anand, P., Puppe, C. and Pattanaik, P. K. (eds), *Oxford Handbook of Rational and Social Choice*. Oxford: Oxford University Press.

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

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

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

McKenna, M. (1997). Alternative Possibilities and the Failure of the Counterexample Strategy. *Journal of Social Philosophy* 28: 71–85.

Ostrum, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Otsuka, M. (1998). Incompatibalism and the Avoidability of Blame. *Ethics* 108: 685–701.

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

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.

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.

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: 57–80.

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

Widerker, D. (2000). Frankfurt's Attack on the Principle of Alternative Possibilities: A Second Look. *Philosophical Perspectives* 14: 181–201.

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.

Wyma, K. (1997). Moral Responsibility and Leeway for Action. *American Philosophical Quarterly* 34: 57–70.

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