Degrees of Causation

Matthew Braham
Faculty of Philosophy, University of Groningen, The Netherlands
m.braham@rug.nl

Martin van Hees
Faculty of Philosophy, University of Groningen, The Netherlands
Martin.van.Hees@rug.nl

(April 15, 2008)

Abstract The primary aim of this paper is to analyze the concept of degrees of causal contribution for actual events and examine the way in which it can be formally defined. This should go some way to filling out a gap in the legal and philosophical literature on causation. By adopting the conception of a cause as a necessary element of a sufficient set (the so-called NESS test) we show that the concept of degrees of causation can be given clear and even empirical meaning. We then apply a game theoretical framework to derive a measure of causal contribution. Our favoured measure turns out to be a generalised version of the normalized Penrose-Banzhaf index of voting power.

1. Introduction

There is an important and vexing issue in legal and moral theory that has yet to be clarified in a satisfactory way, that of assigning *degrees of causation* to persons whose actions played a role in bringing about some outcome. That is, given all the information relevant to a particular circumstance and outcome, what is the meaning of the statement 'Smith made a larger causal contribution to the realization of some state of affairs than Jones'? This paper is aimed at providing such a clarification.

One can of course immediately ask what, if any, meaning can be attached to the notion of degrees of causation for particular or singular events (or 'cause in fact' in legal terminology).[1] For, if some event is an antecedent condition for a particular state of affairs, then that event is simply a cause or in complex situations a causal factor, and not more or less of a cause than some other antecedent condition. Like being pregnant, a causal ascription is categorical and thus has no magnitude.

[1]For the reader unfamilar with the literature on causation, 'singular causation' refers to those causal relations in which the relata are particulars, i.e. a relation that does not exemplify a pattern captured by a 'covering law'. The term 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.

Yet, the fact remains that not only do we make statements about the degrees of causation in ordinary speech but in many legal and moral contexts the concept is required for the practical purpose of distributing responsibility.[2] This is especially the case if we take into account justice considerations in which punishment and reward are supposed to be at least in line with, if not proportional to, a person's contribution to an outcome; and it is also the case for institutional design where we have to assign duties and obligations based upon institutionally determined capacities for action.[3]

The contexts in which we need to make intelligible statements about degrees of causation are not uncommon and are particularly prominent in cases of collective action. These include those circumstances in which different parties (acting either alone or in concert) perform actions which together contribute to the emergence of a state of affairs but in which that outcome is not divisible ('non-severable' in legal terminology) in terms of those actions (i.e. aspects of the resulting state of affairs cannot be uniquely traced back to particular actions);[4] and they occur in cases of causal overdetermination or those circumstances in which the causes or causal factors of a state of affairs are cumulatively more than sufficient to generate the state of affairs and which may be asymmetric in character.

An example of a non-divisible (non-severable) outcome is the case discussed by Wright (1985) in which different firms cause a harm by simultaneously emitting amounts of effluent into a river and in which no single firm could cause the damage alone. Another version of this pollution problem is if the different firms emit the same amounts of effluent but of different kinds. One firm emits two different toxins into the river and a second firm only one, but the emission of three types form the *post factum* necessary and sufficient conditions for the harm. An example of overdetermination in which we may want to impute different degrees of a causal impact to the agents involved is when two firms simultaneously pour toxins into a river with one firm dumping twice as much as the other, but in which the actions of both firms are in itself sufficient to cause a certain harm.

The common characteristic which all these examples exhibit is that in each there is an asymmetry in the actions available to, and performed by, the agents. This suggests that causal efficacy is not an all-or-nothing affair but a difference in degree. How are we to capture this asymmetry in our causal judgements? Even in relatively simple cases where there is no overdetermination, such as in the first two pollution examples above, it is not obvious whether the causal contributions to the harm should be taken to be equal or not.

---

[2]For examples from law, see Payne (1955), Hart and Honoré (1959, 232), and Moore (1999, 13). For a more general statement see Feinberg (1968). However, in none of this literature is the concept given any qualification.

[3]This is particularly important for what Miller (2001) calls 'remedial responsibility'.

[4]If an identifiable part of a harm has been found to have been caused by one act alone, such a part is termed 'severable'. See Hart and Honoré (1959, 225ff).

Given the practical demand, can any factual content be given to the notion of degrees of causation? One prominent answer, which is the one that Hart and Honoré (1959, 233ff) provided in their seminal monograph *Causation in the Law* is to say that the concept, while valid, is inescapably vague with its substance being provided by attributive terms of ordinary language. That is, 'degrees of causation' is captured by locutions such as the 'chief' or 'main' or 'principal' cause, or of 'more important', 'effective' or 'potent' causes. Lifting from Hart and Honoré, we say, for instance, 'His failure in the examination was due more to his not working than the difficulty of the papers' or 'The main [chief, principal] cause of [factor in] his success as a miler was his assiduous training.'

Critics of the concept of degrees of causation would be justified in retorting that what Hart and Honoré have to say is vacuous because it says nothing about *why* one causal factor is more important, effective, potent etc. than another. Ordinary language only teaches us how the concept of degrees of causation is commonly used but does not determine the truth-value of such quantitative attributions.

To determine the truth-value of quantitative attributions of causal impact we need to settle two issues, neither of which has been tackled in the literature. One is the *units of measurement* that are assignable to the respective actions that brought about an outcome; the other is the *method of aggregating* such units. With few exceptions,[5] philosophers have focussed exclusively on the meaning of the claim that an agent's action was 'causal' or a 'causal factor' for an outcome and not on the meaning of the claim that one person's action had, in some sense, a greater impact than another's. This has left a significant void in our understanding of causality in general and responsibility in particular. The current situation is like saying that Smith is taller than Jones yet at the same time confessing that we have neither a scale of measurement nor a general understanding of the meaning of the expression 'is taller than'.[6]

Can this gap be filled? Can we render meaningful quantitative attributions of causal impact that distinguish between the various gradations of 'more than' in a non-arbitrary way? We believe the answer is a qualified 'yes' and will develop the

[5]Recent and notable exceptions are Lewis (2000, 2004), Chockler and Halpern (2005), and Vallentyne (2008). However, in all three cases the idea that causal contribution can come in degrees is assumed rather than justified. Maslen (2004), in particular, criticises Lewis for this. But as we will briefly note later (fn.14), Lewis did, albeit implicitly, suggest the outlines for a justification. Chockler and Halpern also suggest a metric of degrees of causation (although they call it a measure of responsibility it is for all intents and purposes only based on attributions of causal contribution) but justify it only on the grounds that thinking of causality in 'all-or-nothing' terms 'does not at times allow us to make the distinctions that we may want to make' (p.93). However, as with Lewis, the outlines for a justification is implicit in their] of a causal contribution.

[6]This problem has been recognized for a long time in law and is the reason why many legal theorists argue that tort liability should be based on 'fault' or the 'gravity of harm' and not on the degree of causal contribution (Payne (1955), Hart and Honoré (1959)). The problem addressed here is similar to the problem of making quantitative statements of a person's 'freedom' (see Carter (1999) and van Hees (2000).)

argument in six steps. Our first task (Section 2) is to outline the concept of causal contribution that we will be working with. Here we will simply take on board what is increasingly becoming accepted – at least in the legal literature – as the most plausible and comprehensive account of *actual causation*, the so-called 'NESS test'. We present two versions of it – a weak and a strong one. Next (Section 3), and this is where our contribution gets off the ground, we will defend the idea that causal contribution has extensive magnitude by drawing a distinction between *specific* and *overall* causal impact of an action. While the former does not come in degrees, the latter does because, as we will demonstrate, overall causal impact is the aggregation of specific impacts as determined by the NESS test. In the third step (Section 4) we recast the two versions of the NESS test in the language of game theory and use this in our fourth step to construct and examine two corresponding and very simple functions for measuring overall causal impact (Sections 5 and 6). Finally we show how our analysis is related to existing measurements of individual power in voting games (Section 7). We conclude (Section 8) with some remarks about responsibility.

2. Unravelling Causal Ascriptions

We need to start with some preliminaries about the conception of causality that we will adopt. First, and very broadly, we take a 'cause' to be a relation between distinct *events* (in our case, *actions*) in the same time series, in which one ancestral event, C, has the efficacy to produce, or be part of the production, of another, the effect, E, i.e. C is a condition for E. Second, we will assume, as is generally the case in legal theory, and particularly in tort law, that causation is a relation of dependency to be understood in terms of necessary or sufficient conditions (Honoré 1995). This means we will set aside the definition of a cause or causal condition as being a 'probability raising event', i.e. C is a cause of E if the occurrence of C raised the probability that E would occur.[7] While this is certainly a route to follow, it has a fundamental weakness that Lewis (2004, 79) has noted: not all probability raising events should count as a cause or causal condition because it is possible to raise the probability of an outcome via some unactualized event.[8]

There are different ways of saying that C produced or contributed to the production of E. One way is to say that C is *necessary* for E: if E occurs then C must have preceded it. Another way to attribute causal status to C is to say that C is *sufficient* for E: if C occurs then E follows. Another way is to say that C was *necessary and sufficient* for E. Yet another way is to say a cause is conditional

[7] A recent example is Vallentyne (2008). In the legal literature, the probabilistic approach has been developed by Rizzo and Arnold (1980). See also Kaye and Aickn (1984).

[8] Lewis's example is the planting of two indeterministic bombs A and B on two different aeroplanes. Bomb A bombs goes off while B does not. A newspaper runs the headline 'Airline Bomb Disaster'. While bomb B raised the probability of the headline, it certainly did not cause it.

dependency that satisfies what is known as the 'but for test': C is a cause of E if and only if '*but for* the occurrence of C, E would not (have) occur(red).[9]

As is well known, necessity, sufficiency, and but-for dependence turn out to be unworkable. Sufficiency and the but-for test can fail if an outcome required a complex of events, such as collective action; and necessity (as well as but-for dependency) will fail whenever there is causal overdetermination. To elaborate, 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. No individual's action of lifting is sufficient, none is necessary, and none satisfies the 'but-for' condition. Hence nobody can be ascribed causal status for the rescue. A clearly unsatisfactory judgement. There are many other germane examples in the literature but we need not rehearse them here.[10]

The increasingly accepted solution to this conundrum is to use a form of dependence that subordinates the necessity criterion to the one of sufficiency and replaces the idea of identifying C as '*the* cause' ('sufficiency' or 'necessity and sufficiency') or '*a* cause' ('necessity') of E with that of identifying C as a 'causally relevant factor' of E. This conception ascribes C causal status for E if it satisfies the following criterion known as the NESS test (the mnemonic is given by the bold letters) (Wright 1988, 1020):

Definition 2.1 (Weak NESS test) *C is a causal condition for E if and only if (i) C obtains, (ii) E obtains, and (iii) C is a **n**ecessary **e**lement of a **s**ufficient **s**et of conditions for E.*[11]

Here the 'but for' test is nested: 'but for the presence of C the particular set of conditions containing C would not be sufficient for E'. In Pearl and Halpern's (2005) language, a NESS test is a test of 'contingent dependency' which, roughly speaking, says that C is a cause of E if E depends on C under *some* contingency that was present on the occasion.[12]

---

[9] Although in the legal literature the 'but-for' test is often taken to be a test of necessity (Honoré 1995), strictly speaking this is mistaken because the counterfactual formulation of the but-for test entails that E should be absent in the nearest possible world in which C is not the case rather than in *some* possible world (nearest or not). Even if we take the necessary and sufficiency conditions for singular causation as *post factum* necessity and sufficiency, as is often the case in legal theory and as we shall also do (on the importance of this, see Mackie (1965)), a but-for condition need not be a necessary condition.

[10] For an overview, see Braham (2008). See Halpern and Pearl (2005).

[11] 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). The NESS test was actually first stated in Hart and Honoré (1959) and can be traced back to J.S. Mill.

[12] Halpern and Pearl (2005) provide a fully fledged formal structure of the NESS test that takes

The NESS test as formulated is not the test as it is commonly used. A restriction that is generally made and accepted is 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.[13] This yields what we call the 'strong' NESS test because it has a fourth condition:

Definition 2.2 (Strong NESS test) *C is a causal condition for E if and only if (i) C obtains, (ii) E obtains, and (iii) C is a necessary element of a sufficient set of conditions for E, and (iv) all other elements of that sufficient set are necessary as well.*

There are a number of important and general remarks to be made with respect to either version of the NESS test. The first is that because we are dealing with singular causation we assume that it is clear what the relevant events are that are to be taken into the picture of the particular case at hand. This means, for instance, that 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. To establish, say, the causal impact of Smith's driving on the accident in which he was involved, we assume that we can ignore the bad temper of Smith's partner at breakfast even though that temper caused Smith to drive with less attention than he otherwise does. For our purposes we can make the additional simplifying assumption that the relevant facts about the world and the mechanisms are known to us. So, to return to our earlier example, when we say that one of the walkers makes a causal contribution to the rescue of the injured jogger we assume that we know which factors determine who can lift up the fallen tree. It is much like saying that not only do we know the outcome of a vote (who or how many voted for each alternative) but also we know the procedure and decision rule that was used (which alternative won).

Second, the generality of the NESS test means that it can collapse into the canonical dependencies of necessity and sufficiency: if C is a member of every sufficient condition of E for a given instance of E then C is necessary for E; and if C is sufficient for one sufficient condition, then C is sufficient for E.

Third, inclusiveness of the NESS test means that it easily accounts for cases of overdetermination because 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, consider the injured jogger again. What the NESS test does is to resolve the excess sufficient set (the set of actions of 'lifting the tree trunk') into its component sufficient conditions and to check if an action is a necessary part thereof. Assuming, for the sake of simplicity, that each of the rescuers lifts with equal force, there are three possible sets of actions

---

into account some of the problems that the NESS test faces. We do not, however, need the apparatus here.

[13] See Mackie (1965,1974), Wright (1988), and Halpern and Pearl (2005)

that are minimally sufficient for the rescue and each rescuer belongs to at least one (in fact two) of these. Consequently each of the rescuers' actions can be attributed causal status even though none of those actions was itself a necessary, sufficient, or but-for condition for the rescue.

Regarding the difference between the weak and strong NESS test, the strengthening of the original NESS test appears to be eminently sensible because it prunes the extraneous events in the sufficient condition (i.e. those events that are devoid of any causal efficacy for the outcome such as a child picking up twigs in the injured jogger example). Marc-Wogau (1962), for instance, gives an ordinary language justification which makes use of this intuition. He argues that irrelevant events are not intended when we speak of the elements of a 'sufficient condition'. However, the intuitive sensibility seems to be the sole justification for the use of the strong version. The only other consideration that we have encountered is the one that Honoré (1995, 365) attributes to Mackie (1965, 1974). According to Honoré, for Mackie it is *ideally* the case that the sufficient condition is minimal because this is the way that we discover causal regularities. Mackie was, however, quite aware that statements of singular causation need not imply any such generalization. The lack of a proper justification of the use of the strong rather than the weak NESS test is all the more interesting because it turns out to be anything but innocuous when it comes to deriving an index of causal contribution. In fact, as we shall show, the measurement of degrees of causal contribution should be formulated in terms of the weak rather than in terms of the strong and standard version of the NESS test.

3. Specific and Overall Causal Contribution

Now that we have a reasonably clear idea of what it means for an event to be ascribed causal status for some outcome, we can move on to the question of giving factual content to 'degrees of causation'. To avoid any linguistic and conceptual confusions we need to make two clarifying statements. First, in line with the NESS test we take this locution to mean 'degrees of causal contribution' or 'impact' or such like. Second, when we speak of events we mean both the actions that have been performed by the agents and the combination of the actions which form a sufficient condition for the outcome. In the injured jogger case, the individual acts of lifting as well as the combination of these acts are events. Next, and this is the burden of this section, in order to defend the idea that causal contribution can come in degrees we need to identify the primitive events that are to be counted. To achieve this we will draw a distinction between *specific* and *overall* causal contributions of an outcome. The former is the primitive notion while the latter is derivative and obtained by aggregating the specific contributions. This aggregation gives us 'degrees of causal contribution'.

When we use the NESS test to ascribe causal status to an action for some

outcome we are making a very specific statement about that action. The test only informs us of the fact that the action was necessary for a condition to be sufficient for that outcome on a specific occasion. Hence the category of a *specific* causal contribution. Since the NESS test only checks whether or not a relation exists between a specific action and a specific outcome, being a causal factor does not have a magnitude if one assumes that 'existence' does not come in degrees.

The categorical nature of NESS conditions comes to the fore if we give it some formal structure. Anticipating the game-theoretic framework that we present later, we do so in set-theoretic terms. Define a set of actions S that have been performed as an event that is a *critically sufficient condition* for E if (i) it is sufficient for E and (ii) there is at least one $s \in S$ such that $S - \{s\}$ is not sufficient for E. S is a *minimally sufficient condition* if (i) it is sufficient for E and (ii) for all $s \in S$, $S - \{s\}$ is not sufficient for E. For instance, assume $\{\underline{a}, \underline{b}\}$ is the only critically (in fact minimally) sufficient condition for E (the criticality of an action is indicated by an underscore). Both of the actions a and b form a NESS for E by dint of them being being elements of $\{a, b\}$ which occurred. However neither of the two actions can be considered as being 'more' or 'less' of a cause than the other because regardless of how 'big' or 'small' a is, a could contribute to the production of E on the specific occasion only if $\{a, b\}$ materialized. To see this, consider, for instance, a committee that uses a weighted voting rule (such as in the EU Council of Ministers) in which each member of the committee has a different number of votes and a proposal is approved if a particular quota of votes is reached. Suppose we have five voters with voting weights 35, 20, 15, 15, and 15 and a quota of 51. Suppose that the first two voters vote in favour of a given proposal and the remaining voters vote against it. The proposal is approved. Despite the fact that in terms of voting weights the first voter can be said to be 'bigger' (in fact 'heavier') than the second, it does not follow that in *this* contingency he is 'more' of a cause than the second voter. Both versions of the NESS test (in fact the but-for test as well) designate both voters as having made a causal contribution, irrespective of their 'size'.

Accepting this straightforward point does not, however, rule out that the causal impact of an action can have magnitude if understood as a derivative concept. Consider the weak NESS test and assume, by way of example, that E and $\{a, b, c\}$ have occurred. Assume that the critically sufficient conditions are $\{\underline{a}, b, c\}$, $\{\underline{a}, \underline{b}\}$ and $\{\underline{a}, \underline{c}\}$. It seems correct to say that a makes 'more' of a contribution to E on the occasion than either b or c. One way of explaining this is to say that this is true in virtue of the fact that a is a necessary (and 'but-for cause') for E while b and c can be said to be merely 'NESS causes' (on the occasion). The 'more' in this scenario refers to the fact that a satisfies a stricter criterion. But this argument does not work in general. Consider the following: E and $\{a, b, c, d\}$ occurred with the critically sufficient sets being $\{\underline{a}, b, c\}, \{\underline{a}, \underline{b}\}, \{\underline{a}, \underline{c}\}, \{\underline{a}, \underline{d}\}, \{\underline{a}, b, d\}, \{\underline{a}, c, d\}$, and $\{\underline{b}, \underline{c}, \underline{d}\}$. Here we only have what can be termed 'NESS causes' and no 'necessary',

'sufficient' or 'but-for' causes, so the strictness condition will fail to generate an answer although it seems reasonable to postulate that a has more causal impact than b, c, or d. Why?

The answer lies with the implicit operation of quantifying over the number of critically sufficient sets for E that each action belongs to. Stated somewhat differently, a appears to make more of a contribution to E because there are more token events that are sufficient causes of E that require the presence of a. In the previous example, a belongs to three such events while the other three actions belong to two. So a statement of the form 'a makes more of a causal contribution than b to the occurrence of E' can be given explicit and numeric sense in this framework because it takes as its primitive units the specific contributions that are picked out by the NESS test and those units can be counted. This gives us the concept of *overall* causal contribution. Note that this quantity can also be said to be 'empirical' in the sense that the contributions are based on individual actions which have occurred and can be observed to have occurred.

This complex quantitative attribute is what allows us to speak of degrees of a causation.[14] To avoid confusion, and to clarify our language in this matter, we do not say that 'a is more of a cause of E than b' but rather 'a has made a larger causal contribution than b' to the production of E. The reason for such a judgement is that there are a greater number of instances in which a forms part of an critically sufficient condition – a token event – for E. Clearly, the same applies if we take the strong instead of the weak NESS test; the only difference here is that we would focus on the number of instances in which an action forms a necessary part of a *minimal* rather than of a *critical* condition.

4. A Game-Theoretic Formulation of the NESS-Tests

The preceding sections have been concerned with supporting a single claim: that it is possible to speak in a qualified way about 'degrees of causation', and because we have identified certain empirical objects that can be counted we can do so in

[14]The idea of quantifying over token events that are ascribed causal status for E is essentially the same idea as was suggested by Lewis (2004, 91). Here Lewis defines causation as *influence*: 'Where C and E are distinct actual events, let us say that C influences E iff there is a substantial range C_1, C_2, ... of different not-too-distant alterations of C (including the actual alteration of C) and there is a range E_1, E_2, ... of alterations of E, at least some of which differ, such that if C_1 had occurred, E_1 would have occurred, and if C_2 had occurred, E_2 would have occurred, and so on. Thus we have a pattern of counterfactual dependence of whether, when, and how upon whether, when, and how.' Lewis then says: 'Influence admits of degree in a rough and multi-dimensional way. How many different C_i's are there?' (p.92). Lewis does not actually elaborate on the nature of the C_i's, but in our structure these are the critically or minimally sufficient conditions. It is worth noting that Maslen's (2004, 592) criticism that Lewis did not provide an answer to the question of the nature of degrees of causation (as 'influence') is not entirely accurate. Although Lewis did not formally define an *influence function* on the space of C_i's, he did indicate in very general terms what the primitive inputs of such a function are.

essentially the same way we do when we speak about 'degrees of red'. We now have to consider how to aggregate these objects. More formally, we have to define a causality function on the space of actions which we take as the primitive events of a sufficient condition (a possibly composite event) that will yield a measure that expresses degrees of causal contribution. To do so, we present the outlines of the game-theoretic framework in terms of which we shall define our measure and give the game-theoretic renditions of the weak and strong forms of the NESS test (Definitions 2.1 and 2.2).

Let X be a set of outcomes with at least two members (i.e. $\#X \geq 2$, where $\#(\cdot)$ denotes a set's cardinality) and $N = \{1, \ldots, n\}$ a set of individuals. $G = (S_1, \ldots, S_n, f)$ is a *game form* (on X and N): each S_i is a set of strategies for each individual i, and f is a function from the set of all strategy combinations, or plays, onto X. Since the function is onto X, each element of X is an outcome in at least one play.

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: $\pi(s_T) = \{f(s_T, s_{N-T}) \mid s_{N-T} \in \Pi_{i \notin T} S_i\}$.

Definition 4.1 *A T-event s_T is a* sufficient condition *for $A \subseteq X$ if and only if $\pi(s_T) \subseteq A$.*

If A is a singleton set, say $A = \{x\}$, we drop the curly brackets and simply say that the event is a sufficient condition for x. Similarly, we shall often write in such cases $\pi(s_T) = x$ rather than $\pi(s_T) = \{x\}$ or $f(s_T) = x$ for the case in which $T = N$.

For any s_U and s_T, call s_U a *subevent* of s_T if $U \subseteq T$ and 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. The critical and minimally sufficient conditions can now be formulated as:

Definition 4.2 *An event s_T is a critically sufficient condition for A if and only if (i) s_T is a sufficient condition for A and (ii) there is at least one $i \in N$ such that the proper subevent $s_{T-\{i\}}$ is not sufficient for A (i is called A-critical for s_T).*

Definition 4.3 *An event s_T is a minimally sufficient condition for A if and only if (i) s_T is a sufficient condition for A and (ii) for all $i \in N$ the proper subevent $s_{T-\{i\}}$ is not sufficient for A.*

We can now define the game-theoretic versions of the weak and strong NESS conditions as follows:

Definition 4.4 (Weak NESS test) *Given a play s_N, an individual strategy s_i is a weak NESS condition for A if, and only if, there is an event $s_T \subseteq s_N$ such that (i) s_T is a critically sufficient condition for A, (ii) i is A-critical for s_T.*

Definition 4.5 (Strong NESS test) *Given a play s_N, an individual strategy s_i is a strong NESS condition for A if, and only if, there is an event $s_T \subseteq s_N$ such that (i) s_T is a minimally sufficient condition for A, (ii) i is a member of T.*

5. Causation Indices for Simple Strategies

There are two basic approaches to fashion a measure describing degrees of causation. The modest approach is to try to derive an *ordinal* comparison of degrees of causation attributed to agents and their actions. The more demanding approach is to define a measure that also allows for *cardinal* comparisons. Given that our ultimate target (although not in this paper) is to be able to say something about legal and moral responsibility and therefore be able to say something about the distribution of punishment, rewards, or burdens, defining a cardinal measure is the approach we will follow. The problem with an ordering is that it generates insufficient information to be an argument in a responsibility function because, all things equal, it will only inform us which agents made a larger causal impact than others but not tell us by how much. That is, the deficiency of an ordinal measure is its insensitivity to the *extent* to which a person participated in bringing about an outcome.

Thus, given a game form G and a play $s_N = (s_1, \ldots, s_n)$ that results in some outcome $\pi(s_N)$, how might such a cardinal value function be defined? Firstly, we want such a function to assign a value between 0 and 1 for each action such that if the action was necessary and sufficient on the occasion it takes a value of 1 and, at the other extreme, if it was not a NESS condition then it takes a value of 0. Secondly, we would want the function to be an index that measures the relative share of the total causal condition, i.e. the sum of values is equal to 1.[15]

Before deriving our measure we need to make one further distinction between two types of games, because this will have a significant implication for the measures. The distinction is the one between situations in which each strategy describes one specific action and situations in which a strategy may stand for a combination of actions that can be carried out simultaneously. In the first case we speak about the

[15]Causal contribution is not to be conflated with responsibility. One difference, according to us, is that the responsibility that we attribute need not have this constant-sum property. That is, it may make sense to say a person's actions form a partial share of the cause even though the person is burdened fully with the responsibility. See Zimmerman (1985) and Parfit (1984, 67ff). We discuss this issue in the conclusion.

individuals having *simple strategies* whereas in the second case they are said to have *complex* strategies. As complex strategies pose extra technical difficulties, we will first derive our metric for simple strategies and treat the complex case separately in the next section.

Very generally, we want to obtain a function that expresses the relative frequency of an action being a NESS condition on a specific instance of an outcome.[16] Because of the general acceptance of strong the version of the NESS test, it is natural to focus on the relative frequency by which an action satisfies the strong NESS test.

Given s_N, let $\mathcal{M}_i$ be the set of all sub-events s_U that form a minimal sufficient condition for $\pi(s_N)$ and in which i performs an action:

$$\mathcal{M}_i = \{s_U \mid s_U \subseteq s_N \text{ is a minimally sufficient condition for } \pi(s_N) \text{ and } i \in U\}.$$

The (normalized) relative frequency by which an action is a strong NESS condition is given by:

$$\alpha_i(G, s_N) := \frac{\#\mathcal{M}_i}{\sum_{j \in N} \#\mathcal{M}_j}. \tag{5.1}$$

Despite its simplicity and intuitive appeal, the measure is problematic. It can generate highly unreasonable rankings as the following example shows. The basic problem with the focus on minimally sufficient conditions is that it does not discriminate between the causal contribution made by an action that is sufficient and one that is 'merely' necessary.

Example 5.1 Let $N = \{1, 2, 3, 4, 5\}$ be a five person committee having to make a choice about $X = \{x, y\}$. The voting rule specifies that x is chosen if, and only if, (i) 1 votes for x, or (ii) at least three of the players 2–5 vote for x. Assume all individuals vote for x: $s_N = (x, x, x, x, x)$. Hence, we have a case of overdetermination: 1's vote is a sufficient condition for the realization of x, and so is any combination of the votes of at least three of the other individuals. Voter 1 is a member of only one minimally sufficient condition, i.e. the event consisting only of his action of voting for x. Since each of the other individuals is a member of three minimally sufficient conditions (each individual is a member of three 'minimal' majorities not containing 1) the measure defined in equation (5.1) yields:

[16] There are of course other possibilities than focusing on the frequency with which an action is a NESS condition, but this means departing from the very natural and straightforward procedure that we pursue here. An example is that due to Chockler and Halpern (2004) who present a measure that depends on the minimal number of changes that have to be made before an action becomes a NESS condition. However this measure, which they call a measure of responsibility, does not determine the share of an action in bringing about an outcome but 'the extent to which there are other causes' (p.94).

$$\alpha_1 = \frac{1}{13}, \quad \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \frac{3}{13}.$$

This is counterintuitive. By focusing on minimally sufficient conditions, the measure ignores the fact that anything that players 2–5 can do to achieve x, player 1 can do, and in fact more – he can do it alone. So, where does this leave us? It has to be borne in mind that the justification for α_i lies first and foremost with the NESS condition and then with the minimality restriction. That is, to measure causal contribution we are not a priori wedded to the minimality condition. The obvious solution here is to notch up the causal contributions by taking the weak NESS test as the appropriate rendition of causal factors. That is, for any game form G and a play of it $s_N = (s_1, \ldots, s_n)$ that results in some outcome $\pi(s_N)$, we count the instances in which an action is critical (Definition 4.2) irrespective of the causal contribution of other actions. This gives us a measure analogous to α_i but with the summation taking place over $s_U \in \mathcal{C}_i$ instead of $s_U \in \mathcal{M}_i$, where $\mathcal{C}_i$ is the set of all sub-events s_U that form a critically sufficient condition for $\pi(s_N)$ and which contain i's strategy:

$$\mathcal{C}_i = \{s_U \mid s_U \subseteq s_N \text{ is a CSC for } \pi(s_N) \text{ and } i \text{ is } \pi(s_N)\text{-critical for } s_U.\}.$$

We put,

$$\beta_i(G, s_N) := \frac{\#\mathcal{C}_i}{\sum_{j \in N} \#\mathcal{C}_j}. \tag{5.2}$$

Consider Example 5.1 once again. The new measure now takes into account the full range of sufficient events in which at least one action is necessary: all events in which 1 votes for x and at most two others as well, and all events in which exactly three individuals other than 1 vote for x. This yields:

$$\beta_1 = \frac{11}{23}, \quad \beta_2 = \beta_3 = \beta_4 = \beta_5 = \frac{3}{23}.$$

It is interesting to note that there is a close relation between the two indices that we presented and the indices that have been presented as measures of power in voting bodies. The metric α_i corresponds with the so-called 'Public Good Index' introduced by Holler (1982, 1983), whereas β_i is closely related to the so-called normalized Penrose-Banzhaf index (Felsenthal and Machover 1998). In Section 7 we shall make these relationships more precise and discuss the relation between causation and power. We now turn to the analysis of complex game forms.

6. Causation and Complex Strategies

Irrespective of the precise definition of a value function describing the degree of causal contribution, any such value will have to take into account that individual strategies can be complex in the sense described in the previous section. To see why this is relevant, we examine a number of examples. We start first with the overdetermined case and then show that the same problems emerge even for the more elementary in situations in which all actions are necessary. Consider first:

Example 6.1 Players (say firms) 1, 2 and 3 dumped different toxins in a river. Each of the firms' strategies consisted of a single action: dumping a fixed quantity of the toxin denoted by T_1, T_2, and T_3 respectively. Any of the three actions was in itself sufficient to kill all the fish in the river. Applying the causation index β_i, each firm is allocated a causal contribution of 1/3 to the outcome of the fish being dead (the result is the same if we use the strong NESS index, α_i).

The allocation is to be expected: it is a case of overdetermination in which each of the companies has contributed 'equally' to the resulting outcome. It is the same as the classic over-determination case in which two assassins simultaneously shoot and kill their victim. Each act of shooting is sufficient for the death of the victim, and each act thus is a NESS condition. However, to tease out the effect of complex strategies, consider the following variation of the above example:

Example 6.2 Firm 3 transfers its activities to Firm 1 (e.g. Firm 1 buys out the business of Firm 3) so that Firm 3 ceases to dump T_3 in the river. Firm 1 now dumps toxins T_1 and T_3 into the river and Firm 2 dumps T_2. All else remains as in Example 6.1. Both the strategy of player 1 (dumping T_1 as well as T_3) and the strategy of player 2 (dumping T_2) form NESS conditions for the resulting outcome. The causation indices α_i and β_i allocate equal shares of the cause for the dead fish – each will have a value of 1/2 (on either index).

The allocation here is problematic. Firstly, there has been no change in the type and quantity of toxins that have been dumped in the river, and secondly there has been no change in the number of events that have brought about the pollution of the river and the death of the fish. The only feature that has changed is *who* performed the actions of dumping. Whereas in Example 6.1 Firm 1 could only cause the death of the fish with a single action, it can now do it with two because Firm 3's action has shifted to Firm 1. The reason for the counterintuitive result is that the measures we presented only focuses on strategies as such and thereby ignore the components of those strategies. By not discriminating between the differences in causal contributions made by those components we lose important information.

If we want to distinguish the various actions that different strategies comprise, we need to have a richer informational framework. For this purpose we introduce

the notion of an *inner game form* of a play $s_N = (s_1, \ldots, s_n)$ of a game form G. The inner game form will be used by us to distill information about the causal impact of the various actions constituting a person's complex strategy, information that cannot be obtained from the original game form. Once we have that information we use it to obtain a more adequate value of the causal contributions made by the various individuals in the original game form.

Basically, an inner game form is a game form in which each of the actions that is part of some person's strategy in s_N forms a separate strategy. Let $m = \Sigma_{i \in N} \# s_i$ and $M = \{1, \ldots, m\}$. Treating actions played by different individuals as different from each other, m is the total number of different actions performed in s_N. We now define:

Definition 6.1 (Action Assignment) *An action assignment is any n-tuple* $\theta = (\theta_1, \ldots, \theta_n)$ *such that:*

1. *For all* i, θ_i *is a mapping from* s_i *in the set* $M = \{1, \ldots, m\}$*;*
2. *For all* i, j *and all* $a \in s_i$, $b \in s_j$ ($a \neq b$ in case $i = j$)*:* $\theta_i(a) \neq \theta_j(b)$.

An action assignment thus assigns exactly one member of M to one action being played in s_N. We now define an inner game form as follows.

Definition 6.2 (Inner Game Form) *Given a game form* $G = (S_1, \ldots, S_n, f)$ *and a play* s_N *of it (with outcome denoted by* x*), an inner game form is a game form* $G^* = (S_1^*, \ldots, S_m^*, f^*)$ *on* M *and* A ($A \subseteq X$) *such that:*

1. *There is some action-assignment* θ *and some* t *not being an element of any* s_i *such that for all* $i \in N$ *and all* $a \in s_i$, $S^*_{\theta_i(a)} = \{a, t\}$*;*
2. $f^*(s_1^*, \ldots, s_m^*) = x$ *if* $s_i^* \neq t$ *for all* $i \in M$*;*
3. *For all plays* $s_m^*, \bar{s}_m^*$ *of the inner game form:*
   if $\{i \in M \mid s_i^* = t\} \subseteq \{i \in M \mid \bar{s}_i^* = t\}$*, then* $f^*(s_m^*) \neq x$ *implies* $f^*(\bar{s}_m^*) \neq x$.

Each individual in an inner game form has only two strategies, one that corresponds with an action played in the play s_N of the original game form and one consisting of an action, t, that was not played at all in s_N (Clause 1). In case each action performed in s_N is played in the inner game form as well, the outcome of the play of the inner game form is the same as that of s_N (Clause 2). If some play s_M^* leads to an outcome different from x, then any other play that only differs because some members who did not play t in s_M^* now do so, will not lead to x either (Clause 3). The rationale for this condition is to exclude 'perverse' strategies such as Firm 1 in Examples 6.1 and 6.2 simultaneously dumping a toxin *and* an antidote into the river. The clause thus precludes the existence of 'preventive' actions. Note that

since any game form has at least two distinct outcomes, the clause also entails that the play in which all individuals play t will have an outcome different from x.

Though different inner game forms can be associated with a particular play s_N of a standard game form G, we assume that there is exactly one 'appropriate' one. To illustrate the definition, consider again the game form described by Example 6.2.

Example 6.2 continued If we let T_0 denote the action of not dumping any toxin, the relevant inner game form is like a three-person game form in which one player can play T_0 or T_1, one can play T_0 or T_2, and one can play T_0 or T_3, i.e. in the inner game form each action of each strategy that was played in the original game form now forms a separate strategy of a 'separate player'. By assumption, a play of the inner game form will lead to the outcome in which the fish are killed if, and only if, at least one of the players adopts a strategy other than T_0.

Given some G and s_N associated with an inner game form G^*, we can now define the *refined causation indices* α_i^* and β_i^*. Let $a_M = (a_1, \ldots, a_m)$ denote the play of G^* in which each action played in s_N is played as well, that is, for all $i \in M$, $a_i \neq t$. We then define:

$$\alpha_i^*(G, s_N) = \sum_{a \in s_i} \alpha_{\theta_i(a)}(G^*, a_M) \quad \text{and} \quad \beta_i^*(G, s_N) = \sum_{a \in s_i} \beta_{\theta_i(a)}(G^*, a_M) \qquad (6.1)$$

The refined causation indices are thus obtained by simply adding up in the inner game form the (non-refined) causation values of the players corresponding to the various components of the original strategies. Obviously, in case each of the strategies in the play s_N is a simple rather than a complex strategy (consists of only one action), the refined causation values coincide with the non-refined ones. The refined indices can therefore be seen as generalizations of the non-refined ones.

Example 6.2 continued We now first calculate the non-refined causation values for the three individuals in the play (T_1, T_2, T_3) of the inner game form. On both the α_i and the β_i versions, the value is 1/3 for each player. Summing up for each player in N the causation values of the actions performed by him yields the refined values:

$$\alpha_1^* = \beta_1^* = \frac{2}{3} \quad \text{and} \quad \alpha_2^* = \beta_2^* = \frac{1}{3}.$$

Now consider a more complicated instance of overdetermination.

Example 6.3 Firm 1 dumped toxin T_1 and T_3 into the river while Firm 2 dumped toxin T_2. The minimal conditions for the death of the fish now are (i) T_1 and (ii)

T_2 and T_3 together. Assume all toxins are dumped. Since there are two minimal sufficient conditions in the corresponding inner game form (the events (T_1) and (T_2, T_3)), the shares in the cause for the death of the fish on the strong NESS test are:

$$\alpha_1^* = \frac{1}{3} + \frac{1}{3} = \frac{2}{3} \quad \text{and} \quad \alpha_2^* = \frac{1}{3}.$$

Taking the weak NESS test yields a different allocation since there are three critical sufficient conditions (the minimal one plus (T_1, T_2) and (T_1, T_3)) in which T_1 is the critical strategy. As there is only one critical sufficient condition containing both T_2 and T_3 (viz. (T_2, T_3)) we get:

$$\beta_1^* = \frac{3}{5} + \frac{1}{5} = \frac{4}{5} \quad \text{and} \quad \beta_2^* = \frac{1}{5}.$$

Thus far we have considered only inner game forms in cases of overdetermination. They are also important if there is no overdetermination:

Example 6.4 Firm 1 dumped toxin T_1 and T_3 into the river while Firm 2 dumped toxin T_2. The three actions form together the necessary and sufficient condition for the outcome. Using the non-refined values, yields a causal impact of 1/2 for each player. If we take the inner game form however, we derive the same conclusion as with the refined values in Example 6.2: $\alpha_1^* = \beta_1^* = 2/3$, and $\alpha_2^* = \beta_2^* = 1/3$.

The various examples that we used thus far concern the combination of qualitatively different types of actions. We assumed that it is indeed possible to make such qualitative distinctions, that is, we assumed that we can associate a unique inner game form to the original game form. Sometimes, however, the actions performed by individuals differ only quantitatively, say when the firms in our various examples submit different quantities of the *same* toxin. Inner game forms are helpful here as well. By assuming that we can treat the emission of certain quantities of toxin (or the exertion of a certain degree of strength when one lifts a tree trunk, or the intake of a certain quantity of wine, etc.) *as if* they constitute different actions. To see how this works, consider the following example:

Example 6.5 Firm 1 dumps 2 liters of toxin T in the river, whereas Firm 2 dumps 0.5 liter of the same toxin. The *threshold value* for killing all fish is the emission of 1 liter of T into the river (an emission of 1 liter is the necessary and sufficient condition for the fish being killed). To yield the inner game form we assume that each emission of k liters of toxin T forms a separate action, where k is the greatest common divisor of the various quantities of T dumped by the agents and of the threshold value. Thus k will equal 0.5. The corresponding inner game form yields

$$\alpha_1^* = \beta_1^* = \frac{16}{20} = \frac{4}{5} \quad \text{and} \quad \alpha_2^* = \beta_2^* = \frac{4}{20} = \frac{1}{5}.$$

Note that k being a common divisor entails that the α^* and β^* values will always coincide if there is a single type of 'quantifiable' action involved.[17]

The surprising feature of the last example as well as of Example 6.3 is that despite Firm 2's powerlessness to either unilaterally pollute the river or to prevent the river from being polluted, Firm 2 still has a share of the causal contribution. The examples reveal that a person may be causally efficacious for an outcome without the possibility to force or prevent that outcome against the will of others.[18] Any sense of paradox can, however, be quickly dismissed because we can easily find an explanation in a standard Lewis type counterfactual analysis. In the case of Example 6.3, when we determine the efficacy of the actions in the subevent (T_2, T_3) we do so by comparing it to a near-by possible world in which the only difference to the actual world is one in which Firm 1 only dumped Toxin 3. In this case, Firm 2's dumping of Toxin 2 is a necessary condition for the death of the fish in that world, despite the fact that Firm 1 could unilaterally pollute the river and kill the fish.

7. Causation and Power

Examples 6.3 and 6.5 raised the issue of power and its relationship to causation. The examples shows that making a causal contribution does not necessarily imply *preventive* power. This does not contradict the intuitively close relationship between 'power' (as we have briefly defined it above) and 'causation'. [19] In fact in this section we shall show that the relationship holds for an important class of games, so-called 'simple games', which are used to model collective decision making in committees.

A simple game is a game in cooperative form, which means that it allows for the formation of binding agreements between the players, and divides subsets of players $S \subseteq N$, 'coalitions', into 'winners' and 'losers' (this is the meaning of 'simple'). Although the game forms that we have used are non-cooperative, we show how these can be seen to 'induce' a simple game. We then show that the causal contribution index based on the strong NESS test, α_i, corresponds to Holler's power-index. The weak NESS test index, β_i, is shown to correspond to the (normalized) Penrose-Banzhaf power-index.[20]

---

[17] Additionally it is useful to remark that it is not necessary to use the greatest common divisor as the value for k. The use of any other common divisor will yield the same allocation vector.

[18] This is the classic definition of power due to Max Weber (in Gerth and Wright Mills 1948, 180). For an overview of definitions of power see Braham (2008).

[19] The relationship between power and causation is analyzed more generally in Braham (2008) where it is demonstrated that a power ascription is a subset of a causal ascription (the result is based on the work of both current authors).

[20] Readers familiar with the literature on power might ask why we did not formulate a causation index that corresponds with the Shapley-Shubik power index (Shapley and Shubik 1954). This

Definition 7.1 *A simple game (on N) is an ordered pair $(N, \mathcal{W})$ with $\mathcal{W}$ being a set of subsets of N (the set of 'winning' coalitions) satisfying: (i) $\emptyset \notin \mathcal{W}$, and (ii) for all $S \in \mathcal{W}$ and all $T \supset S$, $T \in \mathcal{W}$ (monotonicity).*

Let $\mathcal{G}$ be the set of all simple games. A power index is a function that assigns for each game in $\mathcal{G}$ a vector of real numbers each component of which represents the power of a player in the game $(N, \mathcal{W})$. Given a simple game $(N, \mathcal{W})$, let $\mathcal{M}_i(\mathcal{W})$ be the set of all minimal winning coalitions in $\mathcal{W}$ which contains i, i.e. $\mathcal{M}_i(\mathcal{W}) = \{S \in \mathcal{W} \mid i \in S$ and there is no $T \in \mathcal{W}$ for which $T \subsetneq S\}$.

Definition 7.2 *The Holler power index is the function h_i defined as:*

$$h_i(N, \mathcal{W}) := \frac{\#\mathcal{M}_i(\mathcal{W})}{\sum_{j \in N} \#\mathcal{M}_j(\mathcal{W})}. \tag{7.1}$$

The (normalized) Penrose-Banzhaf index is not defined in terms of the number of minimal winning coalitions an individual is a member of, but in terms of the number of times the individual is 'critical': that is, how often it can turn a winning coalition into a losing one. Let $\mathcal{C}_i(W)$ be the set of all winning coalitions in $\mathcal{W}$ in which i is critical, i.e. $\mathcal{C}_i(\mathcal{W}) = \{S \in \mathcal{W} \mid \ S - \{i\} \notin \mathcal{W}\}$.

Definition 7.3 *The Penrose-Banzhaf power index is the function b_i defined as:*

$$b_i(N, \mathcal{W}) := \frac{\#\mathcal{C}_i(\mathcal{W})}{\sum_{j \in N} \#C_j(W)}. \tag{7.2}$$

There is an obvious formal similarity between h_i and the α indices and between b_i and the β indices. We now establish more precisely the relation between a power

---

is without doubt a possibility, and Chockler and Halpern (2004) have recently suggested it for developing an index of responsibility. The trouble, however, is that not all of the axioms that Shapley (1953) used to define the index can be justified. In particular, the axiom of 'additivity', which concerns the effect of the union of two distinct games has been criticized (Felsenthal and Machover 1998, 195) in the power context and it seems seems even more problematic for the case of singular causation. After all, the union of two games yields an entirely different structure of causal relations and there is no a priori reason to suppose that the causation values in that structure will be a function of the two situations from which it has been derived. Moreover, we note that a similar argument entails that an objection raised against the normalized Penrose-Banzhaf index, viz. its violation of the so-called 'transfer axiom' (Felsenthal and Machover 1998, 251ff), loses its force when its applied to the measure β_i. The transfer axiom concerns the effects of shifts of *weights* of players in a weighted simple game. In our framework such a shift of weights amounts to a change of the effects of the *actions* performed by individuals, and hence to a substantive change of the causal relations. This is not equivalent to transferring actions from one player to another as in the case of Examples 6.1 and 6.2. For that reason, the transfer axiom loses its intuitive force as a adequacy criterion for a measure of singular causation.

index and causation index by transforming a game form and a play of it into a simple game. We do so by (i) viewing each action that has been performed in the game form as an individual player and (ii) stipulating that a subevent of the play s_N is a sufficient condition for $\pi(s_N)$ if, and only if, the set of individuals corresponding to those actions is winning in the simple game.

Let G^* be an inner game form (on M and A) associated with a game form G (on N and X) and a play s_N of G. Let θ be the action assignment that yielded G^*. For any $T \subseteq M$, let $s_{\theta^{-1}(T)}$ denote the unique event consisting of those combinations of individual actions the image of which under θ equals T. Finally, let $(M, \mathcal{W})$ be the simple game in which $T \in \mathcal{W}$ if and only if (i) T is a non-empty subset of M and (ii) $s_{\theta^{-1}}(T)$ is a sufficient condition for $\pi(s_N)$. We then have:

Proposition 7.1 *For any G and s_N:*

$$\alpha_i^*(G, s_N) = \sum_{a \in s_i} h_{\theta(a)}(M, \mathcal{W}) \quad \text{and} \quad \beta_i^*(G, s_N) = \sum_{a \in s_i} h_{\theta(a)}(M, \mathcal{W}).$$

Proof. The proof is straightforward.

For any game form and any play s_N of it in which individuals only have simple strategies, the relation between causation and power is even more obvious. Since each strategy s_i that is played contains only one element, we have $M = N$. Taking an action assignment that assigns i to the component of strategy s_i, we obtain the following corollary of Proposition 1:

Corollary 7.1 *For all plays s_N of G consisting of simple strategies only:*

$$\alpha_i^*(G, s_N) = h_i(N, \mathcal{W}) \quad \text{and} \quad \beta_i^*(G, s_N) = b_i(N, \mathcal{W}).$$

The relation between causal contribution and power thus established can be summarized as: the overall causal contribution made by an individual equals the sum of the power of the actions performed by that individual. Or even more informally: the causal contribution of an individual equals the power of its actions.

8. Conclusion

Earlier in this essay (section 5) we stated that the ultimate aim of studying degrees of causation is to be able to say something about the attribution of responsibility. We want to wind up with a few remarks on the implication of our analysis for our understanding of responsibility.

First, whichever causation index one would favour – and we favour the β indices for the reasons given – the values themselves say nothing in general about the degree

of responsibility. More often than not, responsibility is a matter of culpability, which is commonly seen as a qualitative rather than quantitative attribute. The common law, for instance, divides guilty felons into four categories: 'perpetrators', 'abettors', 'inciters' (all of these are 'accomplices') and 'criminal protectors' (Feinberg 1968, 684). Thus the problem of assessing degrees of responsibility for an outcome is not only a matter of assessing the extent of each individual's causal contribution to that outcome but also involves an assessment and integration of various other dimensions such as degrees of initiative, degrees of authority, the gains from the activities involved, and perhaps most difficult of all, the degree of voluntariness.

Consider, for example, the case in which an order is issued to a group of soldiers. The military code stipulates that in the event of a conflict between orders issued by their superiors the soldiers must (and it is assumed that they will) obey the highest ranked officer. As it happens there is no conflict: the senior officer who is present on the occasion remains silent; it is only the subordinate officer who issued the command. Given that both the issuing of the command by the subordinate officer and the silence on the part of the senior officer are NESS-conditions, and in fact the only ones, our measures will allocate each officer equal shares of the causal impact. If the outcome is the wrongful execution of a prisoner of war, then despite the equal causal shares, the senior officer is clearly more responsible given the authority that he had to prevent the execution. Consequently, a causation index will only be useful for distinguishing between individuals in cases where the culpability of the actions are alike or where responsibility and causal contribution are taken to be synonomous.[21] Example 6.4 is the illustrative case. If emitting toxins into a river is a culpable act, then, *ceteris paribus*, Firm 1 is more responsible than Firm 2 because it performed two culpable acts each of which was a weak NESS condition for the death of the fish compared to Firm 1's single act of performing a weak NESS condition. Note that this result required the innovation of the inner game form. Otherwise both firms would be attributed equal responsibility.

Second, another feature of responsibility that the derivation of the causation indices brings to light is that it tests the power-responsibility relationship. Although it is common practice to associate responsibility with power, this is mistaken in general. Peter Morriss (Morriss 1987, 39), for instance, asserts that 'The connection between power and responsibility is, then, essentially negative: you can deny all responsibility by demonstrating lack of power.' That is, you can prove your innocence for the outcome by being able to demonstrate that you could not have prevented it. '[P]ower', says Morriss, 'is a necessary (but not sufficient) condition for blame: if you didn't have the power, you are blameless.'[22] Obviously, if an agent had the power to force the particular outcome in question, such as Firm 1 in Example 6.3,

[21]The most natural application of a causality function is in tort law. However, legal scholars are very divided on whether liability should be apportioned on the basis of causal contribution or comparative fault. See Moore (1999).

[22]There are many other examples, see Connolly (1983), Reeve (1982), and Holler (2007).

that agent will be responsible if the action is culpable by some standard. However, the fact that power may imply responsibility does not mean that powerlessness to prevent can be used to exonerate agents. If preventive power is a necessary condition for responsibility, then Firm 2 in Example 6.3 could not be held to account for the pollution of the river. This appears to be wrong, given that there is – even if it is relatively weak – a causal connection between Firm 2's action and the pollution of the river and the death of the fish.

Third, there is a significant methodological by-product that has been thrown up. The fact that certain voting power indices can be shown to be a special case of causation indices suggests an additional use and interpretation for these measures. Note that in the case of simple games there is a perfect overlap between causal contribution and power. This permits us to apply the normalized Penrose-Banzhaf index as the appropriate means for allocating responsibility for the outcome of committee votes. Although Holler (2007) has recently made a similar proposal and suggests the use of his index for this purpose, our justification fundamentally differs from his. Holler argues that we should use power to attribute responsibility, which we reject; and Holler suggests the appropriate way to measure power is to use the strong NESS test, which we also reject. Nonetheless, what Holler and ourselves agree upon, and for which we believe that we have delivered the correct theoretical foundations, is the idea that power indices can be used for more than designing fair voting systems raises a whole host of interesting theoretical questions and possibilities of new normative analyses.

Last, we have thus far only discussed issues involved in relating singular causal contribution to *retrospective* responsibility. Responsibility is however not just about what has happened, but also about *prospective* events and outcomes. What remains open is the derivation of a measure of causation that captures what we can do, not just what *we have done*. This is vital to our understanding of our positive duties to others.

References

Braham, M. (2008). Social power and social causation: Towards a formal synthesis. In M. Braham and F. Steffen (Eds.), *Power, Freedom, and Voting*. Springer.

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

Chockler, H. and J. Y. Halpern (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research 22*, 93–115.

Connolly, W. E. (1983). *The Terms of Political Discourse*. Oxford University Press.

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

Felsenthal, D. S. and M. Machover (1998). *The Measurement of Voting Power*. Edward Elgar.

Gerth, H. and C. Wright Mills (1948). *From Max Weber: Essays in Sociology*. Routledge.

Halpern, J. Y. and J. Pearl (2005). Causes and explanations: A structural-model approach. part i: Causes. *British Journal of Philosophy of Science 56*, 843–887.

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

Holler, M. J. (1982). Forming coalitions and measuring voting power. *Political Studies 30*, 262–271.

Holler, M. J. (2007). Freedom of choice, power, and the responsibility of decision makers. In A. Marciano and J.-M. Josselin (Eds.), *Democracy, Freedom and Coercion: A Law and Economics Approach*. Edward Elgar.

Holler, M. J. and E. W. Packel (1983). Power, luck and the right index. *Zeitschrift für Nationalökonomie (Journal of Economics) 43*, 21–29.

Honoré, A. M. (1995). Necessary and sufficient conditions in tort law. In D. Owen (Ed.), *Philosophical Foundations of Tort Law*, pp. 363–385. Oxford University Press.

Kaye, D. and M. Aickin (1984). A comment on causal apportionment. *Journal of Legal Studies 13*, 191–208.

Lewis, D. (2000). Causation as influence. *Journal of Philosophy 97*, 182–197.

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

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.

Marc-Wogau, K. (1962). On historical explanation. *Theoria 28*, 213–233.

Maslen, C. (2004). Degrees of influence and the problem of pre-emption. *Australasian Journal of Philosophy 82*, 577–594.

Miller, D. (2001). Distributing responsibilities. *Journal of Political Philosophy 9*(4), 453–471.

Moore, M. S. (1999). Causation and responsibility. In E. Frankel, F. D. Miller, and J. Paul (Eds.), *Responsibility*, pp. 1–51. Cambridge University Press.

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

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Payne, D. (1955). Reduction of damages for contributory negligence. *Modern Law Review 18*, 344–355.

Reeve, A. (1982). Power without responsibility. *Political Studies 30*, 77–86.

Rizzo, M. and F. Arnold (1980). Causal apportionment in the law of torts. *Columbia Law Review 80*, 1399–1429.

Shapley, L. S. (1953). A value for n-person games. In H. Kuhn and A. W. Tucker (Eds.), *Contributions to the Theory of Games*. Princeton University Press.

Shapley, L. S. and M. Shubik (1954). A method for evaluating the distribution of power in a committee system. *American Political Science Review 48*, 782–792.

Vallentyne, P. (2008). Brute luck and responsibility. *Politics, Philosophy and Economics (forthcoming)*.

van Hees, M. (2000). *Legal Reductionism and Freedom*. Kluwer.

Wright, R. (1985). Causation in tort law. *California Law Review 73*, 1735–1828.

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.

Acknowledgements

Forerunners of this paper have been discussed at seminars at the universities of Hamburg, Groningen, and Warwick. We are grateful to comments received at these seminars. Special thanks go to Dan Felsenthal, Manfred Holler, Erik Krabbe, Theo Kuipers, Moshé Machover. Allard Tamminga should be singled out for his detailed and insightful written comments on a draft of this paper.