Cause and Context

Alexander Kaiserman

Jesus College

A thesis submitted for the degree of Doctor of Philosophy

University of Oxford

Trinity Term 2016

For Jaya

Abstract

This thesis comprises an introduction and six papers on causation, freedom and responsibility. Though mostly self-standing, the papers are unified by two common goals – to recognise and analyse the role of *context* in the semantics of causal claims and ascriptions of freedom; and to put metaphysical approaches to causation into closer contact with actual causal reasoning in science and the law.

Chapter One defends a contextualist semantics of causal language that combines the ancient idea that causes necessitate their effects with Angelika Kratzer's semantics of modality. Chapter Two extends this approach to ascriptions of freedom, by combining Kratzer's account with the principle that an agent acts freely only if she could have acted otherwise. Chapter Three explores a neglected view which combines David Lewis's counterfactual account of causation with his counterpart-theoretic approach to *de re* modality. Chapter Four proposes an amendment to the interventionist account of causation in response to a worry raised by John Campbell about causation in psychology. Chapter Five motivates the idea that causation is a relation to which multiple events can contribute to different degrees, and defends a novel account of an event's degree of contribution to a causing of an effect. Chapter Six then argues, from a conception of tort law as a system of corrective justice, that a defendant should be held liable for a claimant's losses only to the degree to which the defendant's wrongdoing contributed to the causing of the claimant's harm.

Acknowledgements

I would like to thank everyone who has taken the time to engage with the ideas in this thesis, especially Simona Aimar, Sara Bernstein, Harjit Bhogal, John Campbell, Shamik Dasgupta, Delia Graff Fara, Rachel Fraser, Johann Frick, Anandi Hattiangadi, Thimo Heisenberg, Richard Holton, Jonathan Ichikawa, Mark Johnston, Boris Kment, Adam Lovett, Neil McDonnell, Peter Millican, Laurie Paul, Caleb Pearl, Jonathan Schaffer, Jason Schukraft, Josh Shepherd, Jan Sprenger, Sandy Steel, Henning Strandin, Victor Tadros, Andy Yu, my fantastic supervisors Ofra Magidor and John Hawthorne, four anonymous referees for *Australasian Journal of Philosophy*, *Philosophical Studies* and *Erkenntnis*, and audiences in Aarhus, Austin, Cambridge, Dartmouth, Edinburgh, Lille, Liverpool, London, Lund, New York, Oxford, Princeton, San Diego, Stockholm, Vancouver and Warwick. I am very fortunate to have benefited from their comments, suggestions, ideas and words of encouragement over the past three years.

I was able to pursue my graduate studies only thanks to generous funding from Jesus College, the Oxford philosophy faculty, the Arts and Humanities Research Council and the Henry Fund, whose Procter Fellowship enabled me to spend a year of my studies as a visiting student in Princeton – I am extremely grateful to them for their support.

Thanks last, but not least, to my parents, Paul and Catherine, for all their encouragement over the (many) years it has taken me to get this far. I hope to have convinced them that philosophy is about more than asking questions; even if I know better than to suggest that I have conclusively answered any.

Table of Contents

Introduction

The six papers in this thesis are intended to be self-standing, and each can be read and understood independently of the others. But they are also unified by a number of common themes. In this introduction, I will provide a brief summary of the chapters below, explaining how they relate to each other and to the broader philosophical literature.

1. Causation and Freedom in Context

Knowing, it seems, requires eliminating all possibilities of error. Yet possibilities of error are everywhere – I believe that I have hands, for example, but I'm apparently unable to rule out the possibility that I am a handless brain in a vat. Doesn't it thereby follow that I know next to nothing?

Acting freely, it seems, requires the possibility of acting otherwise. Yet constraints on action are everywhere – my making coffee this morning, for example, was metaphysically necessitated by the initial state of the universe and the laws of nature, at least if those laws are deterministic. Doesn't it thereby follow that no action is free?

Causing an effect, it seems, requires bringing it about "of necessity" (Aristotle, *Metaphysics* IX.5).[1] Yet potential interfering influences are everywhere – the flame wouldn't have followed the striking of the match had all the oxygen been sucked out of the room, for example. Doesn't it thereby follow that no individual event causes any other?

In response to the first 'paradox', David Lewis (1996) defends a well-known *contextualist* semantics of knowledge ascriptions. On his view, ascriptions of knowledge are relativized to a set of *proper presuppositions* (or equivalently, to a set of possible worlds that we are

[1] Quoted in Anscombe (1971: 134).

not *properly ignoring*), which is determined by – and varies with – the conversational context. When we do epistemology, we often find ourselves in contexts relative to which we know nothing much at all; but most everyday ascriptions of knowledge are made in more permissive contexts, in which more is being properly presupposed, relative to which there are many more things that we know. In this way, we can trace a path "between the rock of fallibilism and the whirlpool of scepticism" (Lewis 1996: 550): We account for the Moorean intuition that many actual ascriptions of knowledge are *true*, without giving up on the idea that to utter 'I know that *p*' and 'I haven't eliminated a possibility in which not-*p*' – *in the same context* – is to contradict oneself.

In Chapters One and Two, I defend similar responses to the other two 'paradoxes' above. My strategy is to develop a contextualist semantics of ascriptions of causation and freedom by combining two familiar principles – the principle that causes *necessitate* their effects, and the principle that an agent acts freely only if she *could have acted otherwise* – with a mechanism developed by the linguist Angelika Kratzer (1977, 1981) for modelling the context-sensitivity of modality, according to which restricted domains of possible worlds are picked out by two conversational backgrounds, the *modal base* and the *ordering source*. Not only can this strategy dissolve the 'paradoxes' above, it also helps to explain a number of otherwise puzzling features of the way we make and evaluate claims about causation and freedom, in everyday life, but also in science and the law.

2. Causes and Counterparts

Suppose I sculpt a statue out of a lump of clay. The lump of clay, it seems, could survive being squashed; the statue, on the other hand, could not survive being squashed. Doesn't it thereby follow that there are *two* objects – the statue and the lump of clay – occupying exactly the same region of space?

Suppose John says 'Hello' to Fred. He says it rather too loudly. But it wasn't John's saying 'Hello' *loudly* that caused Fred to greet him in return; rather, it was John's saying 'Hello'

that caused Fred to greet him in return. (For suppose John's saying 'Hello' loudly hadn't occurred – then John would still have said 'Hello', only not so loudly, and so Fred would still have greeted John in return.) Doesn't it thereby follow that there are *two* greetings – John's saying 'Hello' and his saying 'Hello' loudly – occurring at exactly the same time?

These two 'paradoxes' share a common structure. Yet few have attempted to find a common solution. In Chapter Three, I show that it follows from Lewis's counterfactual account of causation and his counterpart-theoretic approach to *de re* modality that causal claims are relativized to a set of counterpart relations. This suggests a way of accounting for the differing semantic roles of 'John's saying "Hello"' and 'John's saying "Hello" loudly' in modal and causal contexts, without giving up on the intuitive idea that they pick out one and the same event. Though I don't ultimately endorse this solution, I argue that it is better motivated, *by his own lights*, than the view Lewis actually held, and also better motivated than a similar view (currently much in vogue) which relativizes causal claims to sets of 'contrast events'.

3. Causes and Interventions

Lewis's long-term project, pursued over a series of papers (Lewis 1973, 1979, 1986), was to define causation in terms of counterfactuals and to define counterfactuals in terms of a 'closeness' metric on possible worlds. But there is a difficulty for this approach: It's notoriously difficult to define a closeness metric that gets the right results, without implicitly appealing to causal concepts (Edgington 2011). Chapter Four is about an approach to causation which embraces this 'circularity'. In science, causal hypotheses are usually tested by performing controlled manipulations informed by prior causal knowledge. But whilst every account of causation will, of course, find some connection between causal relationships and correlations under experimentally idealized manipulations, Woodward's (2003) *interventionist* account of causation is unique in taking such correlations to be *constitutive* of causation, and not simply evidence for it.

Roughly speaking, Woodward defines causation in terms of correlations under *interventions*, and then defines the notion of an intervention in causal terms. As an analysis of causal claims, then, interventionism is explicitly and unashamedly non-reductive. But it nevertheless purports to be illuminating – in just the sense that functionalism about the mind seeks to better understand mental states in terms of the functional relations between them and not simply by defining each one individually in behavioural terms, interventionism seeks to better understand both causation and experimental manipulation in terms of the connections between them and not simply by defining each of them individually in non-causal terms.

This new approach to the metaphysics of causation has been particularly influential in the philosophy of mind, with some arguing that it has the potential to break the deadlock in traditional debates about causal exclusion and the causal relevance of mental content.[2] But not everyone is so optimistic.[3] In Chapter Four, I examine a worry raised by John Campbell (2007) for the applicability of the interventionist framework to causation in psychology. The problem, he argues, is that it follows from one of the conditions on interventions – the so-called 'surgical constraint' – that an intervention on an intention requires the suspension of the agent's rational autonomy. I argue first that the problem Campbell identifies is in fact an instance of a wider problem for interventionism, extending beyond psychology, which I call the problem of 'abrupt transitions'. I then defend a solution to the problem and explore some of its consequences for causation in psychology.

4. Causal Contribution and Partial Responsibility

Are there 'degrees of causation'? 'Yes', say Moore, Bernstein, and Braham and van Hees:

[2] See Campbell (2010); Raatikainan (2010); Rescorla (2014); Shapiro and Sober (2007); Weslake (forthcoming); Woodward (2008, 2015).

[3] See Baumgartner (2009, 2010, 2013), for example.

> [C]ausation is a scalar relation. Scalar relations are more-or-less affairs, not matters of black or white. (Moore 2009: 275).
>
> [T]here can be comparative degrees of causation, or quantitative differences in causal contribution. (Bernstein forthcoming).
>
> [C]ausal efficacy is not an all-or-nothing affair but a difference in degree. (Braham and van Hees 2009: 324).

'No,' say Pearson, Wright, and Barker and Steele:

> Causation…exists or it does not…one does not speak of "degrees" of causation. (Pearson 1980: 346).
>
> Causation…is not a matter of degree. Some condition either was or was not a cause (in the proper scientific sense) (Wright 1988: 1146).
>
> [T]he idea…[of] relative causal contributions…is seemingly oxymoronic. (Barker and Steele 2015: 67).

In Chapter Five, I motivate a metaphysics of causal contribution that can vindicate both sides of this debate. Causation, I argue, is *not* a scalar relation; it does not come in degrees. But causation relates *pluralities* of events to events, in general; and different causes can *contribute* to a causing of an effect to different extents.

Although philosophers have devoted a great deal of attention to the question of when one event is a cause of another, the question of how to determine its *degree* of contribution to a causing of an effect has been almost completely ignored. Yet the practice of describing one event as 'more of a cause' of an effect than another – and similar talk of 'degrees of contribution', of 'causal potency' or 'causal efficacy', and of 'chief', 'main' or 'principal' causes – is pervasive in many disciplines, including the natural and social sciences, history and the law. In the remainder of this chapter, I defend a novel probabilistic analysis of an event's degree of contribution to a causing of an effect that can make sense of these locutions.

It's widely accepted that you're only responsible for what you cause.[4] But if causal contribution comes in degrees, should we also think of *responsibility* for outcomes as coming in degrees? Chapter Six argues that we should. More specifically, I argue that when a claimant sues a defendant for damages, the defendant should be held liable for the claimant's losses only to the degree to which the defendant's wrongdoing contributed to the causing of the claimant's harm. I show how such an approach can make sense of the fraught legal concept of 'proximate causation' and the mechanisms for 'apportioning' damages in cases with multiple wrongdoers. Though it has some revisionary consequences, I argue that these consequences can be defended, at least on a conception of tort law as a system of *corrective justice*, whose role is to correct injustices inflicted by one person on another, and *not* to redistribute the costs of harms or to punish wrongdoers.

The arguments in Chapter Six are suggestive of further applications to other areas of the law. Talk of 'degrees of causation' has been employed as part of proposed reforms to the mechanism of secondary liability in the criminal law, for example (Moore 2009: 299-303), as well as in recent criticisms of orthodox just war theory (Tadros ms). As I hope to argue to in future work, a rigorous metaphysics of causal contribution can help us better evaluate these proposals too.

5. Two Definitions, One Concept

Hume defined causation twice over.[5] He wrote: "We may define a cause to be *an object followed by another, and where all the objects, similar to the first, are followed by*

[4] Although see Sartorio (2004).

[5] Well, *thrice* over: "The appearance of a cause always conveys the mind, by a customary transition, to the idea of the effect. Of this also we have experience. We may, therefore, suitably to this experience, form another definition of cause; and call it, *an object followed by another, and whose appearance always conveys the thought to that other*." (Hume 1748: 60).

objects similar to the second. Or, in other words, *where, if the first object had not been, the second never had existed*." (Hume 1748: 60).

These are two very different definitions, of course, notwithstanding Hume's apparent insistence to the contrary. The first tries to capture the idea that causes *necessitate* their effects; the second tries to capture the idea that causes are things which *make a difference*. Two quite different traditions have since emerged in the metaphysics of causation literature, each committed to developing one of the two ideas that Hume mistakenly ran together.[6]

I believe that *both* of these ideas are ultimately essential to our concept of causation. Causes, in other words, should be thought of as pluralities of events which *jointly* necessitate, and *individually* make a difference to, their effects. This approach features prominently in the writings of J. L. Mackie (1965, 1974), in particular, and echoes of his views are clearly visible in the papers below. But it should be stressed that at no point in this thesis will I try to turn this rough characterization into a reductive analysis of causal claims – a definition of causation in non-causal terms – nor am I committed to there being such a thing. In particular, I leave open the question of whether the sense in which causes 'make a difference' to their effects is best cashed out in terms of counterfactuals, or interventions, or non-redundancy conditions, or something else entirely. My project is one of "rich characterization" (Sider 2011: 9) as opposed to definition; my aim, to put it another way, is to defend substantive claims about the *structure* of causation and its relations to other concepts, rather than to formulate a list of necessary and sufficient conditions for its instantiation. This may leave some readers disappointed; to them I say, read on – the proof, after all, is in the pudding.

[6] Hall (2004) even argues that there are two concepts of causation – 'production' and 'dependence' – with different extensions.

References

Anscombe, G. E. M. (1971). *Causality and determinism*. Cambridge: Cambridge University Press.

Barker, K. and Steele, J. (2015). Drifting towards proportionate liability: Ethics and pragmatics. *The Cambridge Law Journal 74*(1), 49-77.

Baumgartner, M. (2013). Rendering interventionism and non-reductive physicalism compatible. *Dialectica 67*(1), 1-27.

Baumgartner, M. (2010). Interventionism and epiphenomenalism. *Canadian Journal of Philosophy 40*(3), 359-383.

Baumgartner, M. (2009). Interventionist causal exclusion and non-reductive physicalism. *International Studies in the Philosophy of Science 23*(2), 161-178.

Bernstein, S. (forthcoming). Causal proportions and moral responsibility. In D. Shoemaker (ed.), *Oxford studies in agency and responsibility (vol. 4)*, Oxford: Oxford University Press.

Braham, M. and van Hees, M. (2009). Degrees of causation. *Erkenntnis 71*(3), 323-344.

Campbell, J. (2007). An interventionist approach to causation in psychology. In A. Gopnik and L. Schulz (eds.), *Causal learning: Psychology, philosophy, and computation*, Oxford: Oxford University Press.

Campbell, J. (2010). Control variables and mental causation. *Proceedings of the Aristotelian Society 110*(1), 15-30.

Edgington, D. (2011). Causation first: Why causation is prior to counterfactuals. In C. Hoerl, T. McCormack and S. R. Beck (eds.), *Understanding counterfactuals, understanding causation*, Oxford: Oxford University Press.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, and L. A. Paul (eds.) *Causation and counterfactuals*, Cambridge, MA: The MIT Press.

Hume, D. (1748[2006]). *An enquiry concerning human understanding (ed. T. L. Beauchamp)*. Oxford: Clarendon Press.

Kratzer, A. (1977). What 'must' and 'can' must and can mean. *Linguistics and Philosophy 1*(3), 337-355.

Kratzer, A. (1981). The notional category of modality. In H. J. Eikmeyer and H. Rieser (eds.), *Words, worlds and contexts*, New York: de Gruyter.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs 13*(4), 455-476.

Lewis, D. (1986). Causation. In D. Lewis, *Philosophical papers (vol. 2)*, Oxford: Oxford University Press.

Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy 74*(4), 549-567.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly 2*(4), 245-264.

Mackie, J. L. (1974). *The cement of the universe*. Oxford: Clarendon Press.

Moore, M. S. (2009). *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford: Oxford University Press.

Pearson, R. N. (1979). Apportionment of losses under comparative fault laws – An analysis of the alternatives. *Louisiana Law Review 40*(2), 323-372.

Raatikainen, P. (2010). Causation, exclusion, and the special sciences. *Erkenntnis 73*(3), 349-363.

Rescorla, M. (2014). The causal relevance of content to computation. *Philosophy and Phenomenological Research 88*(1), 173-208.

Sartorio, C. (2004). How to be responsible for something without causing it. *Philosophical Perspectives 18*(1), 315-336.

Shapiro, L. and Sober, E. (2007). Epiphenomenalism – The do's and the don'ts. In P. Machamer and G. Wolters (eds.), *Thinking about causes*, Pittsburgh: University of Pittsburgh Press.

Sider, T. (2011). *Writing the book of the world*. Oxford: Oxford University Press.

Tadros, V. (ms). Causal contributions and liability.

Weslake, B. (forthcoming). Exclusion excluded. *International Studies in the Philosophy of Science.*

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

Woodward, J. (2008). Mental causation and neural mechanisms. In J. Hohwy and J. Kallestrup (eds.), *Being reduced: New essays on reduction, explanation, and causation*, Oxford: Oxford University Press.

Woodward, J. (2015). Interventionism and causal exclusion. *Philosophy and Phenomenological Research 91*(2), 303-347.

Wright, R. W. (1988). Allocating liability among multiple responsible causes: A principled defense of joint and several liability for actual harm and risk exposure. *University of California at Davis Law Review 21*(4), 1141-1211.

Chapter One

Necessary Connections in Context

Abstract: This paper combines the ancient idea that causes *necessitate* their effects with Angelika Kratzer's semantics of modality. On the resulting view, causal claims quantify over restricted domains of possible worlds determined by two contextually-determined parameters. I argue that this view can explain a number of otherwise puzzling features of the way we use and evaluate causal language, including the difference between *causing* an effect and being *a cause* of it, the sensitivity of causal judgements to normative facts, and the semantics of causal disagreements.

1. Introduction

If there is one constant in the history of philosophical investigation into the nature of causation, it's the idea that causes *necessitate* their effects. Anscombe (1971) traces it back to Aristotle's *Metaphysics*. Some early modern philosophers took it to require the *inconceivability* of causes without their effects,[1] a view which naturally leads to a version of occasionalism.[2] But Hume, despite agreeing that "[t]here is a NECESSARY CONNEXION to be taken into consideration" (Hume 1738: 77) between causes and their effects, famously looked elsewhere for the "source of our idea of power or necessity",

[1] "[A]n entire cause, is the aggregate of all the accidents both of the agents how many soever they be, and of the patient, put together; which when they are all supposed to be present, *it cannot be understood* but that the effect is produced at the same instant." (Hobbes 1655: 121-2; my emphasis).

[2] "A true cause as I understand it is one such that the mind perceives a necessary connection between it and its effect. Now the mind perceives a necessary connection only between the will of an infinitely perfect being and its effects. Therefore, it is only God who is the true cause and who truly has the power to move bodies." (Malebranche 1674: 450).

finding it in the "constant conjunction" of similar events and the "determination of the mind to pass from one object to its usual attendant" (Hume 1738: 164-5).

Early critics of Hume's complained that, on his view, "night is the cause of day, and day the cause of night. For no two things have more constantly followed each other since the beginning of the world" (Reid 1788: 249). To account for this, later writers distinguished merely *accidentally* true generalizations from genuine *laws of nature*.[3] This led to what we might call the 'covering law' condition on causation:[4]

> (CL): *C* caused *E* only if *C* and *E* are events of type **C** and **E**, respectively, and it is a law of nature that events of type **C** are invariably followed by events of type **E**.[5]

Not many philosophers still believe that (CL) is true. For one thing, (CL) is inconsistent with the intuition many people claim to have that causation is possible in an indeterministic universe.[6] But even assuming determinism, (CL) leads to what many see as a problematic proliferation of causes. Suppose a short circuit occurs and a fire breaks out shortly afterwards.[7] It's consistent with the laws of nature that a short circuit occurs

---

[3] For example: "There are sequences, as uniform in past experience as any others whatever, which yet we do not regard as cases of causation, but as conjunctions in some sort accidental. Such, to an accurate thinker, is that of day and night…We may define, therefore, the cause of a phenomenon, to be the antecedent, or the concurrence of antecedents, on which it is invariably and *unconditionally* consequent." (Mill 1843: 377; my emphasis).

[4] Davidson (1980: 208) calls this the "Principle of the Nomological Character of Causality".

[5] Note that (CL) is just a necessary condition on causation. Those seeking an analysis of causal claims have typically combined (CL) with other necessary conditions. According to Hume, for example, a cause must also occur prior to, and spatially contiguous with, its effect (Hume 1738: 75-6).

[6] See Mackie (1974: 40-3) for example. I will mostly be assuming here that the fundamental laws of nature are deterministic. I suspect that the insights of this paper could be extended to the indeterministic case, although I will not attempt such a project here. It's worth pointing out, however, that our best candidate for a fundamental physical theory, namely quantum mechanics, is perfectly deterministic (proponents of the so-called 'GRW theory' have responded to perceived conceptual difficulties with quantum mechanics by postulating an *additional* indeterministic dynamical process, called 'collapse'; but there are other deterministic alternatives, such as De-Broglie-Bohm theory, as well as interpretations of unitary quantum mechanics that take it seriously on its own terms, such as the Everett or 'many worlds' interpretation). In any case, making sense of causation in an indeterministic world is a challenge for everyone; the strategy of trying to get deterministic causation right before worrying about the indeterministic case is widely adopted.

[7] This example is adapted from one in Mackie (1965).

without a fire breaking out, because there are no flammable materials nearby, or because there is no oxygen in the atmosphere, or because every atom in the Andromeda galaxy simultaneously quantum-tunnelled through energy barriers 2.5 million years ago and the earth is destroyed by gravity waves before the fire has a chance to develop. Indeed, "for any small region R of space at time t nothing much short of the state of the universe in a sphere with center R and whose radius is one light second (i.e. 186,000 miles) at t – 1 seconds is causally sufficient for determining what will occur…in R" (Loewer 2007: 252). So according to (CL), what caused the fire was really a vast plurality of events,[8] including the short circuit, but also the presence of oxygen, the absence of macroscopic quantum-tunnelling events,[9] and plenty more besides.

Mill embraced this conclusion. "The cause, then, philosophically speaking, is the sum total of the conditions…taken together; the whole of the contingencies of every description, which being realized, the consequent invariably follows." (Mill 1843: 370). From among these things, Mill insists, "we have, philosophically speaking, no right to give the name of cause to one of them, exclusively of the others" (Mill 1843: 366). But the problem is that we *do* routinely distinguish those events which caused an effect from those which were merely nomologically required for the effect to occur. In the case above, for example, it seems very natural in many contexts to describe the short circuit as the

[8] Here, and throughout this thesis, I refer to the relata of causation as *events*. It should be noted that the term 'event' is used in a somewhat idiosyncratic way in the metaphysics literature, however. In natural language, we would ordinarily reserve the term 'event' for *changes* of some kind; the mere presence of oxygen in a room, for example, would not normally be described as an event. But in the philosophy of causation literature it has become common to use the term in a semi-technical way to include 'standing conditions' like the presence of oxygen. Perhaps 'state of affairs' would have been a better term for the category in question, were it not for the fact that it already has a specific meaning in the literature (see Chapter Three, below). Some philosophers have argued that we should admit other entities, such as objects and facts, as causal relata as well as, or instead of, events (Mellor 1995; Stein forthcoming); I will nevertheless assume, with the majority, that all causal claims can be interpreted as predicating relations of (ordered) pairs of events (broadly understood).

[9] I assume here that absence talk picks out ordinary events under negative descriptions, rather than some *sui generis* kind of entity. But see Lewis (1986a) and Schaffer (2005).

cause of the fire and the presence of oxygen as a mere 'background condition'. As Hart and Honoré put it:

> Mill's doctrine…[implies that] every factor necessary for the occurrence of an event is equally entitled to be called 'the cause'. Yet this is not the case: neither the plain man, nor the historian, uses the expression 'cause', or any related expression, in this way. For the contrast of cause with mere conditions is an inseparable feature of all causal thinking, and constitutes as much of the meaning of causal expressions as the implicit reference to generalizations does. (Hart and Honoré 1985: 12).

Some millennia earlier, Plato expressed the same complaint:

> If someone said that without bones and sinews and all such things, I should not be able to do what I decided, he would be right, but surely to say that they are the cause of what I do…is to speak very lazily and carelessly.
>
> To call those things causes is too absurd…Imagine not being able to distinguish the real cause from that without which the cause would not be able to act as a cause. (Plato, *Phaedo 99a-b*).

Some even interpret the proliferation of causes delivered by (CL) as a *reductio* of the very idea that causes necessitate their effects: "To treat causation as a type of necessity, or even as involving necessity, is to take causation as something it is not", according to Mumford and Anjum (2011: 48). I think this is an overreaction. It's well-known that *modal* claims in natural language rarely, if ever, quantify over every nomologically possible world. There are contexts in which I can truly utter the sentence 'Eric can't jump eight metres', even though there is a possible world in which he does jump eight metres (one in which he's on the moon, for example); and there are contexts in which I can truly utter the sentence 'Mary must be home by now', even though there's a possible world in which she isn't home by now. The mechanisms by which the conversational context determines the restricted domains of possible worlds over which these sentences quantify is now fairly well understood. I propose to apply the insights of this literature to

the idea that causes necessitate their effects. My hypothesis is that this approach can explain a number of otherwise puzzling features of the way we use and evaluate causal language, in everyday life, but also in specialist disciplines like science and the law.

The paper is structured as follows. Section 2 briefly introduces Kratzer's semantics of modality. In section 3, I apply Kratzer's framework to the idea that causes necessitate their effects. Sections 4-6 then use the resulting view to explain three neglected features of causal discourse: The distinction between *causing* an effect and being *a cause* of it, the influence of normative considerations on the causal claims we tend to make and endorse, and the semantics of certain kinds of causal disagreements. Section 7 compares my proposal with three alternative approaches, which appeal to pragmatics, ambiguity, and contrastivism respectively. Section 8 concludes.

2. Modality and Context

Modal claims are claims about what can, may, might, must, should, ought to, or has to be the case. It's well-known that such claims can mean different things in different contexts. The modal auxiliary 'must', for example, has a *deontic* reading, as in 'You must pay your taxes', which has to do with rules or prescriptions; a *bouletic* reading, as in 'You must apply for that job', which has to do with an agent's goals or desires; and an *epistemic* reading, as in 'You must be tired', which has to do with what an agent knows. But there is also significant semantic variation even within one of these 'flavours' of modality. Suppose Karin commits murder. There's an obvious sense in which Karin *should* be in jail in light of the law. But there's *also* an obvious sense in which it should be the case, in light of the law, that Karin never committed murder in the first place, and hence isn't in jail. What's going on here?

The success of Kratzer's (1981) solution is its ability to explain all this semantic variation without abandoning the assumption that modal auxiliaries express the same content in every context of use. The basic idea is that modal claims are relativized to a specification

of what we're 'holding fixed' and a method of *ordering* possible worlds. These together pick out a restricted domain of possible worlds – the 'best' worlds, on the relevant ordering, consistent with what we're 'holding fixed'. Something *can* or *may* or *might* be the case exactly if it's the case in at least *one* of these worlds, and something *must* or *should* or *has to* or *ought to* be the case exactly if it is the case in *every* such world. Different readings of modal claims are generated by varying either what we're holding fixed or how we're ordering possible worlds. 'You must give to charity' might be true on a moral ordering but false on a legal ordering, for example. If we hold fixed the fact that Karin committed murder, the legally 'best' worlds are all ones in which Karin is in jail. But if we hold *nothing* fixed, the legally 'best' worlds are all ones in which Karin doesn't commit murder in the first place, and so isn't in jail.

More precisely, modal auxiliaries on Kratzer's view express functions with, not one, but *three* arguments – a proposition (e.g. the proposition that Karin is in jail) and two sets of propositions, the *modal base* **B** and the *ordering source* **O**. Propositions are sets of possible worlds on Kratzer's framework, so **B** is a set of sets of possible worlds. The set $\cap\mathbf{B}$, then, is the set of possible worlds in which all the propositions in **B** are true. **O** then defines a preorder $<_{\mathbf{O}}$ over $\cap\mathbf{B}$ as follows: for all $w, u \in \cap\mathbf{B}$, $w <_{\mathbf{O}} u$ if and only if the set of elements of **O** that are true in w is a proper subset of the set of elements of **O** that are true in u (i.e. $w <_{\mathbf{O}} u$ if and only if $\{p : p \in \mathbf{O} \ \& \ w \in p\} \subset \{p : p \in \mathbf{O} \ \& \ u \in p\}$). Now let max(**B**, **O**) be the set of *maximal* elements of $\cap\mathbf{B}$ on $<_{\mathbf{O}}$ – the worlds such that no world is ranked above them (i.e. $w \in \max(\mathbf{B}, \mathbf{O})$ if and only if $w \in \cap\mathbf{B}$ and there is no $u \in \cap\mathbf{B}$ such that $w <_{\mathbf{O}} u$). These are sometimes called the *accessible* worlds. Something *must* be the case relative to **B** and **O** if and only if it is the case in *every* element of max(**B**, **O**); something *can* be the case relative to **B** and **O** if and only if it is the case in *some* element of max(**B**, **O**); and similarly for other modal locutions.[10]

[10] I'm assuming here that max(**B**, **O**) is always non-empty. This is called the *limit assumption*, and it's endorsed by Stalnaker (1981) and rejected by Lewis (1973) and Kratzer. If the limit

There are different kinds of modal base – sometimes we're interested in what is possible holding fixed all and only those things we know, sometimes what is possible holding fixed certain salient facts about the circumstances. There are different kinds of ordering source – sometimes we're interested in those worlds which satisfy the most legal, moral, stereotypical or social ideals, sometimes those worlds which satisfy the most goals or desires or teleological ends. We can gesture at these parameters explicitly – *given* that Karin has committed murder, then *legally speaking*, she should be in jail. But in most cases we needn't bother, since the conversational context often succeeds in filling in the extra argument places for us. "When we talk to each other, we hardly ever make explicit in view of which circumstances something should be necessary or possible. We may give hints. Usually people understand. And they all understand in pretty much the same way." (Kratzer 1981: 53-4).

Kratzer's theory leaves many questions unanswered [11] and it's not entirely uncontroversial.[12] But it represents the closest we have to orthodoxy about the semantics of modality. It has been enthusiastically applied to a wide variety of expressions, from conditionals (Kratzer 2012) to generics (Krifka *et al.* 1995). But there have been few attempts to apply it to the semantics of causal claims.[13] This is a puzzling lacuna.

---

assumption is false we need to massage the definitions above to avoid trivial results (see Kratzer 1981). I'll ignore these complications from now on for simplicity.

[11] In particular, it cannot by itself explain why some modals have preferred readings (e.g. 'might' likes to be epistemic, 'can' likes to be circumstantial), as well as various phenomena involving the interaction of modality with tense and aspect (Hacquard 2009).

[12] Some think that utterances of sentences like 'Karin should be in jail' don't express anything truth-evaluable, but rather express (in a different sense of 'express') something like the utterer's disapproval of Karin's actions (Horgan and Timmons 2006). Others think that sentences like 'Karin should be in jail' express the same proposition in every context of use, but these propositions are themselves true or false only relative to a *context of assessment* (MacFarlane 2005). Yet others distinguish what is *said* by an utterance from what is semantically expressed by it, and argue that indefinitely many propositions are said by an utterance of 'Karin should be in jail', even though only a single proposition is semantically expressed by it (Cappelen and Lepore 2005). I mention these alternatives merely to set them aside, but it's worth thinking about how my view could be adapted to fit these various approaches.

[13] The only exception of which I'm aware is Szabó and Knobe (2013), who seek to explain certain patterns of empirical data by assuming that participants evaluate target statements by replacing them with 'modal proxies'. My view is in the spirit of Szabó and Knobe's project, although it goes

Although many philosophers have found it plausible that there are deep conceptual connections between causation and modality, philosophical analyses of everyday causal claims have tended to ignore all the rich semantic complexity of modality in natural language; complexity which, if I'm right, carries over into the semantics of causal claims too.

3. Causation and Context

I propose to apply Kratzer's theory to causal claims. On the view I will defend, 'cause' expresses a function (the same function in every context) with, not two, but *four* arguments: A plurality of events $C_1,\ldots, C_n$ (the causes), an event E (the effect), a modal base **B** and an ordering source **O**. As with modal claims, the conversational context is often sufficient to fix **B** and **O** without them having to be explicitly specified (though we can clarify, for example, that the short circuit caused the fire *given that* there was oxygen in the atmosphere).

Now here's a plausible gloss on the idea that causes necessitate their effects:

> NECESSARY CONNECTIONS (NC): $C_1,\ldots, C_n$ collectively caused E only if, *given* that $C_1,\ldots, C_n$ all occurred, E *had* to occur.

In sentences of the form 'Given that φ, it had to be that ψ', the role of the first clause is to add the proposition expressed by φ to the modal base relative to which the second clause is evaluated. In particular, if $C_1,\ldots, C_n$ (un-italicized) are the propositions that $C_1,\ldots, C_n$ (italicized), respectively, occurred, it's true relative to **B** and **O** that E had to occur given that $C_1,\ldots, C_n$ all occurred, if and only if it's true relative to $\mathbf{B}\cup\{C_1,\ldots, C_n\}$ and **O** that E had to occur. Properly spelled out, therefore, (NC) should read as follows:

---

further in that it actually incorporates Kratzer's framework into the *semantics* of causal claims, and applies the resulting theory to a much broader range of phenomena.

(NC): $C_1,\ldots, C_n$ collectively caused E relative to **B** and **O** only if E occurs in every element of max(**B**$\cup\{C_1,\ldots, C_n\}$, **O**).

To determine whether the short circuit caused the fire relative to **B** and **O**, for example, we take the set of possible worlds in which all the elements of **B** are true, *throw out* all those worlds in which the short circuit doesn't occur, order the elements that remain using **O**, and check whether the fire occurs in every maximal element of this ordering – if it doesn't, the short circuit didn't cause the fire relative to **B** and **O**.

I've introduced a hypothesis about the logical forms of causal claims – that they contain two parameters usually determined by the conversational context – and reinterpreted the idea that causes necessitate their effects in light of this hypothesis. Since (NC) is just a necessary condition on causation, I don't purport to have provided an *analysis* of causal claims. Nevertheless, (NC) can explain a number of otherwise puzzling semantic phenomena (sections 4-6) better than the alternatives (section 7). Or so I will argue.

4. Causes and Causing

Consider the following sentence:

(1) Jaya and Fatima lifted the table.

(1) is ambiguous. Read distributively, it follows from (1) that Jaya lifted the table and Fatima also the lifted the table (perhaps at a different time). But the more natural reading is the collective one, according to which Jaya and Fatima lifted the table *together*.[14] These are distinct states of affairs. On the distributive reading the table is lifted twice; on the collective reading it is only lifted once, even though it is lifted by two people. *Lifting*

[14] When I say that Jaya and Fatima lifted the table together, I *don't* mean that what lifted the table was the *set* containing Jaya and Fatima, or the *mereological fusion* of Jaya and Fatima, or indeed any other single thing. In other words, I believe that there is such a thing as irreducibly plural predication. See McKay (2006).

relates pluralities to individuals, in general; and a plurality of people can collectively lift a table without any one of the plurality individually lifting it.

Exactly the same is true of (2):

(2) The driver's drunkenness and the rainstorm caused the car crash.

(2) is also ambiguous. Read distributively, it follows from (2) that the drunkenness caused the crash and the rainstorm *also* caused the crash. On the more natural collective reading, however, the drunkenness and the rainstorm caused the crash *together*. These are distinct states of affairs. On the distributive reading, the crash was caused twice over – it was *overdetermined*, to put it another way[15] – whereas on the collective reading the crash was only caused once, even though it was caused by two events. *Causing* (and, for that matter, *opening*, *authoring*, *preventing*, and so on) relates pluralities to individuals, in general; and a plurality of events can collectively cause an effect without any one of the plurality individually causing it.

The ambiguity of (2) is naturally explained by (NC) – on the distributive reading, the drunkenness and the rainstorm *individually* necessitated the crash, whereas on the collective reading they only *jointly* necessitated it. More precisely, if D is the proposition that the drunkenness occurred and R is the proposition that the rainstorm occurred, (2) is true on its distributive reading relative to **B** and **O** only if the crash occurs in every element of max(**B**∪{D}, **O**) *and* every element of max(**B**∪{R}, **O**); whereas on its collective reading, (2) is true relative to **B** and **O** only if the crash occurs in every element of max(**B**∪{D, R}, **O**).

If a book is collectively authored by multiple people, each of those people was *an author* of the book. To be an author of a book, in other words, is to be one of the people who

[15] Unger (1977) argues that overdetermination in this sense is *incoherent*, since he thinks that the distributive readings of sentences like (2) are contradictory. I don't endorse that view, though Unger mounts a convincing case.

collectively authored it. The same is true for causation. If an effect was collectively caused by a plurality of events, each of those events was *a cause* of the effect. To be a cause of an effect, in other words, is just to be one of the events that collectively caused it. The expressions '*C* is a cause of *E*' and '*C* caused *E*' are therefore *not synonymous*, notwithstanding a widespread tendency among philosophers to use them interchangeably. If *C* caused *E*, then *C* was trivially a cause of *E*. But the converse doesn't follow. On the collective reading of (2), the rainstorm was *a cause* of the crash, but it didn't *cause* it – what caused it was the drunkenness and the rainstorm taken together.

The distinction between *causing* and being *a cause* of an effect is context-sensitive, according to (NC). Suppose, by analogy, that Eric claims that he can jump eight metres. Consider the following exchange between Yena the athlete and David the facetious physicist:

> **Yena**: Eric can't jump eight metres!
>
> **David**: Well technically speaking he *can* – anyone can jump eight metres on the moon, for example.

David's utterance is literally true, of course. But intuitively, it doesn't contradict Yena's utterance because David has changed the context (this is the sense in which David is being facetious). Yena's utterance was true (and David's false) relative to the context in which she made it, because, in particular, there is no accessible world in that context in which Eric is on the moon – we're 'holding fixed', among other things, the fact that Eric is on earth. But everyday utterances are governed by what Lewis calls 'a rule of accommodation': "what you say makes itself true, if at all possible, by creating a context that selects the relevant features so as to make it true" (Lewis 1986c: 251). David's utterance succeeds in changing the context to one relative to which it expresses a truth, by expanding the set of accessible worlds to include worlds in which Eric is on the moon.

Now imagine a similar exchange between David and Ulla the engineer:

Ulla: The short circuit caused the fire.

David: Well technically speaking the short circuit was only one cause among many – the presence of oxygen was also a cause of the fire, for example.

Again, what David says is perfectly true, but it changes the context. Ulla's utterance was true (and David's false) relative to the context in which she made it, because, in particular, there is oxygen in the atmosphere in every accessible world in that context – we're 'holding fixed' the presence of oxygen. But David's utterance succeeds in changing the context to one relative to which it expresses a truth, by expanding the set of accessible worlds to include worlds in which the short circuit occurs without any oxygen in the atmosphere, and hence without the fire. Relative to *this* context, only the short circuit and the presence of oxygen *taken together* necessitated the fire, so that the short circuit was *a cause* of the fire, but didn't cause it.[16]

In an oft-quoted passage, Lewis claimed to be "concerned with the prior question of what it is to be one of the causes (unselectively speaking)" of an effect, and his analysis was "meant to capture a broad and nondiscriminatory concept of causation" (Lewis 1986a: 162). One way of interpreting this comment is to take Lewis to be giving an account of what it is to be a cause of an effect, *relative to every nomologically possible world*; relative, that is, to an empty ordering source and a modal base that only includes the laws of nature. Relative to these parameters, the fire was collectively caused by the short circuit, the presence of oxygen, the absence of macroscopic quantum-tunnelling events, and so on, so that each was *a cause* of the fire, though none of them (individually) caused it. Analysing what it is to be a cause of an effect in *this* sense is a perfectly respectable

[16] It's true that one can felicitously say things like 'The short circuit caused the fire, but of course it didn't cause it alone'. One might be tempted to conclude from this that '*C* caused *E*' and '*C* was a cause of *E*' *are* synonymous, although the former carries a cancellable *presupposition* that *C* was the only cause of *E*. But another option is that the first and the second clause are evaluated relative to different modal bases, either because the context changes mid-sentence, or (more likely) because a single context can supply more than one set of contextual parameters. One finds similar effects in the modal case: On the natural reading of 'John must pay his debts, but of course he can't', the two clauses are evaluated relative to different sets of accessible worlds.

metaphysical project, analogous to analysing what it is for something to be metaphysically possible (i.e. possible, unrestrictedly speaking). But it's not the whole story, at least if what we want is a *semantic* theory of causal claims – a theory of what causal claims (in contexts) *mean*. The fact is that we do routinely – and, it seems, *truly* – make claims about what caused an effect, such as 'The short circuit caused the fire', that fail to mention even a fraction of the vast plurality of events that collectively caused the effect in this unrestricted sense. According to (NC), this isn't down to "principles of invidious discrimination" (Lewis 1986a: 162) or "the inaccuracy of common discourse" (Mill 1843: 370), but rather down to the antecedently accepted and well-understood mechanism by which the conversational context restricts the domains of possible worlds over which certain kinds of expression quantify.

5. Causes and Norms

Consider the following case, from Clarke *et al.* (2015):

> *Collision*
>
> Two cars, one driven by Greta and the other driven by Rachel, were approaching an intersection. Greta had a green light. Rachel had a red light, but she wasn't paying attention. The lights stayed that way. Neither driver stopped, and their cars collided. (Clarke *et al.* 2015: 282).

Now consider the following claims:

(3) Rachel's driving into the intersection caused the collision.

(4) Greta's driving into the intersection caused the collision.

On average, experimental subjects presented with *Collision* agree with (3) and *disagree* with (4), despite the fact that the collision wouldn't have occurred if either Greta hadn't driven into the intersection or Rachel hadn't driven into the intersection.[17] The only

[17] See the results of Clarke *et al.*'s (2015) 'First experiment'.

relevant difference between the two events seems to be a *normative* one: Driving through green lights is what one is *supposed* to do, whereas one *isn't* supposed to drive through red lights. These judgements have been widely replicated in a variety of different settings and appear to be remarkably robust.[18]

(NC) can explain what's going on here. The asymmetry in our causal judgements about *Collision* is due to the fact that the details of the case encourage us to evaluate (3) and (4) relative to something like the following ordering source:

> **O** = {*Everyone approaching a red light stops*, *Everyone approaching a green light carries on*}

Kratzer would call this a 'stereotypical' ordering source – red lights and green lights are associated with certain stereotypical states of affairs, namely stopping and continuing respectively.

Let G be the proposition that Greta continues into the intersection and let R be the proposition that Rachel continues into the intersection. Given a suitable modal base **B**, the elements of max(**B**, **O**) – the 'top-ranked' worlds of this ordering – are all ones in which Rachel stops at the intersection, Greta carries on, and the collision doesn't occur. The elements of max(**B**∪{G}, **O**), therefore, are also worlds in which Rachel stops at the intersection, Greta carries on, and the collision doesn't occur. Hence it's inconsistent with (NC) that Greta's continuing into the intersection caused the collision, relative to **B** and **O** – the collision didn't *have* to occur, even given that Greta's continuing into the intersection occurred. The elements of max(**B**∪{R}, **O**), by contrast, are worlds in which Rachel continues into the intersection, as does Greta, and so the collision does occur.

[18] Clarke *et al.* re-ran their experiment with four different target statements and four different experimental primes, each specifically designed to urge the participants to respect the counterfactual symmetry of the case in their causal assessments. "No matter how plain we made it…that what Greta did made a difference to whether the outcome occurred, participants tended to take Rachel but not Greta to be one of the causes." (Clarke *et al.* 2015: 283). See also Hitchcock and Knobe (2009), Knobe (2010), and references therein.

Assuming the other necessary conditions for causation are satisfied, (NC) implies that Rachel's continuing into the intersection caused the collision, relative to **B** and **O** – given that her continuing into the intersection occurred, the collision *had* to occur, in the relevant sense. Hence in this context, (3) is true and (4) is false. *That's* why subjects agree with (3) and disagree with (4).

Of course, we can consider the same case relative to different contexts. Relative to an *empty* ordering source (and the same modal base), for example, neither Greta's nor Rachel's driving into the intersection individually caused the collision. Rather, the collision was collectively caused by both events, so that Greta's driving into the intersection was *a cause* of the collision, as was Rachel's. (NC) therefore predicts that we will elicit different responses from subjects if we deliberately undermine the salience of the stereotypical ordering source suggested by the case. For example, Clarke *et al.* found that when subjects were instructed to think about what caused the collision "considering the physics of the situation" (Clarke *et al.* 2015: 288), they were much more likely to agree that both events were causes of the collision.[19]

The phenomenon exemplified by *Collision* is pervasive in causal discourse, in everyday life, but also in science and the law. Consider, for example, the legal case of *McKew v. Holland*,[20] in which an employee of the defendant company sustained an injury in a workplace accident. As a result, his left leg was weakened and would frequently become numb. Three weeks after the accident, the claimant lost his balance while descending a steep staircase without a handrail. He fell and sustained a further injury. Neither injury would have occurred but for the company's negligence. But although the company accepted liability for the first injury, it appealed against the decision to hold it liable for

[19] "[F]ocusing participants' attention in a certain way – emphasizing the physics of a situation – can increase the frequency with which they give egalitarian responses." (Clarke *et al.* 2015: 289). This was the only experimental prime that had any significant effect on results.

[20] [1970] SC 20 (HL).

the second, on the grounds that its negligence didn't *cause* the second injury. The House of Lords agreed:

> If a man is injured in such a way that his leg may give way at any moment he must act reasonably and carefully. It is quite possible that in spite of all reasonable care his leg may give way in circumstances such that as a result he sustains further injury. Then that second injury was caused by his disability which in turn was caused by the defendant's fault. But if the injured man acts unreasonably…what follows must be regarded as caused by his own conduct and not the defendant's fault.[21]

Lord Reid seems to be suggesting here that the causal history of the second injury depends on the *normative* question of whether the plaintiff's actions were 'reasonable'. This, of course, is hard to explain on most philosophical accounts of causation. So what's going on here?

There is a strong tradition in legal theory, beginning with the American legal realists (Green 1927, 1929; Edgarton 1924) and recently revived by Wright (1985, 2001, 2011), Stapleton (2001, 2008, 2010) and others, of criticising this kind of causal reasoning in the law for erroneously conflating two distinct stages of legal inquiry. The aim of the first stage should be to establish whether the defendant's wrongdoing was a 'factual cause' of the claimant's harm. This is "a purely factual, non-normative" question (Wright 2007: 289), one which should remain "untainted by normative controversies" (Stapleton 2008: 474). Various tests have been proposed for 'factual causation', most notably the 'but-for test', according to which, *ceteris paribus*,[22] *C* is a 'factual cause' of *E* if and only if *E* wouldn't have occurred but for the occurrence of *C*. So-conceived, a defendant's action

[21] *Id.*, at 25.

[22] The but-for test fails, of course, in well-known cases of redundant causation. In response, Wright (1985, 2011) defends the more sophisticated 'NESS test', according to which *C* is a cause of *E* if and only if *C* is one of a plurality of events $X_1,\ldots, X_n$, and the occurrence of *C* is necessary for the sufficiency of $X_1,\ldots, X_n$ for *E* (where $X_1,\ldots, X_n$ are 'sufficient' for *E* if and only if it is a consequence of a 'causal law' that if $X_1,\ldots, X_n$ all occur, *E* will occur). For comments and criticisms, see Fischer (2006), Fumerton and Kress (2001) and Stapleton (2008). See also Chapter Five, Section 3, below.

will be a 'factual cause' of many, many more injuries than it is reasonable to hold him liable for. "The law's concern that a defendant not be held liable for the infinite stream of consequences flowing from tortious conduct" therefore requires "the limitation of every obligation to a finite set of consequences" (Stapleton 2001: 984). Hence we need a second stage of inquiry, the aim of which should be to establish whether the claimant's harm was "within the scope of the defendant's liability" (Lunney and Oliphant 2013: 119). This is a normative, non-causal question, which may turn on such considerations as "the costs of legal rules and their administration…the dignity of the law…the interest in individual freedom…the relative wrongfulness of different actors…whether allowance of recovery…would be likely to open up the way to fraudulent claims", and so on (Stapleton 2001: 985-6). About the case above, for example, these theorists would argue that the defendant's negligence was a 'factual cause' of *both* the first and the second injury, since but for the defendant's negligence, neither injury would have occurred. The controversy over whether to hold the defendant liable for the second injury, therefore, could only be about whether the second injury was within the defendant's 'scope of liability'. Insofar as Lord Reid presents the controversy in causal terms, then, he is simply confused.[23]

Advocates of this two-stage strategy have recently convinced the American Law Institute to distinguish 'factual causation' from 'scope of liability' in the Restatement (Third) of Torts,[24] "after having confused students, lawyers and courts for decades by its failure to do so in the first and second Restatements", as Wright (2011: 306) sees it. But results elsewhere have been mixed. A recent review of the law of negligence in Australia, for example, approved of such a separation in principle,[25] but nevertheless chose – "[b]izarrely", in Stapleton's opinion – "to retain the umbrella term of 'causation' to

[23] "The courts…confusingly merge the scientific issue of causal contribution with the normative issue of the appropriate extent of legal responsibility for the consequences of one's (legally relevant) conduct." (Wright 2011: 294).

[24] American Law Institute, *Restatement (Third) of Torts: Liability for Physical and Emotional Harm* (2009).

[25] Commonwealth of Australia, *Review of the Law of Negligence: Final Report* (2002) (Recommendation 29).

signify the amalgam of both issues" (Stapleton 2010: 471). And despite repeated calls from legal scholars "to distinguish the normative and context-specific purposive considerations [from] the causation issue" (Wright 2011: 305), Lord Hoffmann insists that "no judge in fact adopts such a two-stage test" (Hoffmann 2011: 4), and goes on to question the right of "academic writers" to impose this "philosophically privileged form of causation" upon them (Hoffmann 2011: 5). As he drily notes: "One side or the other must be missing something" (Hoffmann 2011: 4).

I think (NC) can provide a better explanation of what is going on here. Lord Reid is not conflating two distinct stages of inquiry in the passage above. Instead, he is evaluating the relevant causal claims relative to a *bouletic* ordering source, one according to which the negligence caused the injury only if the claimant *had* to go down the stairs, in the sense of there being *no reasonable alternative* to his doing so in light of his legitimate interests. If there *was* such a reasonable alternative, then the negligence didn't cause the second injury relative to this context, according to (NC), since even given that the negligence occurred, the second injury didn't *have* to occur.

These kinds of bouletic uses of modal terms appear frequently in causal reasoning in the law. In a similar case[26] wherein the claimant, fearing for his safety, jumped from the defendant's coach and broke his leg, Lord Ellenborough argued that the defendant's negligence caused the injury only if the negligence placed the claimant "in such a situation that he *must* adopt a perilous alternative", thereby creating "a *necessity* for what he did".[27] The defendant's failure to fix a defective train door in *Adams v. The Lancashire and Yorkshire Railway Company* was found not to have caused the claimant's injury sustained while trying to close it, since the claimant didn't *have* to try so hard to shut the door: "[I]t was not even *necessary* to do so to avoid inconvenience".[28]

[26] *Jones v. Boyce* [1816] Stark 1 493.

[27] *Id.*, at 495-6; emphasis added.

[28] [1869] LR 4 CP 739, 741; emphasis added.

The judges in these cases are aware, of course, that there are possible worlds in which the defendant's wrongdoing occurs and the claimant's harm does not; the 'necessity' that concerns them is a more restricted, bouletic necessity, and it is relative to this restricted domain of possible worlds that causal claims about these cases are being evaluated.

This analysis casts doubt on the alleged importance of teasing apart the 'normative' and 'factual' elements of causal reasoning in the law. As Stapleton herself notes, "'causation' is a term we use to express diverse information about the world" (Stapleton 2008: 432). We *can* use it to denote a relation which holds between a plurality of events and an effect only if there is no nomologically possible world in which the former all occur without the latter. But this is clearly not the relation of interest to the law. The right meta-normative question to ask is *not*, 'What is the correct scope of liability for consequences of tortious conduct?' – the 'scope of liability' concept is theoretically redundant, and obscures the role of context in the semantics of causal language.[29] What we should ask instead is, 'Relative to which modal base and ordering source does 'cause' express that relation which is required to hold between defendant and claimant for the former to be held legally liable for the latter's harm?'. This is a big question, of course, and I won't attempt to answer it here (although the cases above suggest that the ordering source in question will partly be a function of the claimant's reasonable interests to bodily security – c.f. Chapter Two, Section 6, below). The important point is that lawyers are *not* erroneously conflating two distinct questions when they draw on considerations of reasonableness in causal inquiries. What they are doing is evaluating causal claims relative to non-empty ordering sources; and Wright and Stapleton do not succeed, in my view, in establishing that there is anything theoretically problematic about this.[30]

[29] C.f. Hart and Honoré (1985: 13): We must "chart in some detail the actual use made…in ordinary discourse of the key expressions like 'cause'"; and, in particular, "the way in which a shifting context affects the force and meaning of these expressions".

[30] Stapleton also thinks there are good *pragmatic* reasons to separate 'factual causation' from 'scope of liability'. Since "for many of us the notion of causation has a factual ring", Stapleton worries that "some judges will be tempted to present their [causal] determinations…without adequate normative justification" (Stapleton 2010: 470-1). I think this problem is somewhat

'Factual causation', according to Wright (2011: 305-6), is a "purely scientific matter". But the irony is that scientific causal claims are also often evaluated relative to non-empty ordering sources. Here's a well-known example from molecular biology.[31] The output of the process of protein synthesis in cells counterfactually depends, not only on the specificities of the cell's DNA, but also on the specificities of a whole host of other molecules, including tRNA molecules, the carriers of amino acids. Although "everyone knows that many different molecules and cellular structures play necessary roles in the *in vivo* syntheses" of proteins, however, "biologists and philosophers often talk as if DNA 'produces'" them, with the other molecules merely mediating this causal influence (Waters 2007: 553).

Again, these distinctions are difficult to explain on standard philosophical theories of causation. And again, there is no shortage of people lining up to criticise biological causal reasoning for being "laden with unjustified privileging" (Oyama 2000: S332). One of the central tenets of 'Developmental Systems Theory', for example, is the so-called 'Causal Parity Thesis', according to which causal explanations in biology should "take the whole developmental matrix into consideration" (Faye 2014: 152), rather than focusing on DNA. But despite repeated calls for a "stake-in-the-heart" to "the notion that some influences are more equal than others" (Oyama 1985: 27), the idea that "genetic factors are…privileged" (Franklin-Hall forthcoming) in biological explanations remains as stubborn as ever.[32]

overstated, however – as the passage from *McKew* above illustrates, judges are usually well aware that the causal claims they are evaluating turn on normative questions.

[31] For a good introduction to this debate, see Stegmann (2012).

[32] For alternative attempts to make sense of this phenomenon in causal terms, see Waters (2007), Franklin-Hall (forthcoming) and Woodward (2010). See Weber (forthcoming) for criticisms of these accounts.

(NC) can explain this phenomenon, on the assumption that biologists are often interested in what caused an effect relative to *teleological* ordering sources. As Weber points out:

> Substituting tRNAs…for molecules with different specificities…is not consistent with continuing biological functioning of the process of protein synthesis. For if the cell suddenly contains different tRNAs…this will affect the sequence of other protein molecules made by the cell, which makes it impossible for it to survive. (Weber forthcoming).

In other words, the cell *has* to have the particular tRNA molecules it in fact has to ensure continued proper functioning. A certain DNA sequence therefore necessitates the production of a certain kind of protein, relative to those teleologically top-ranked possible worlds in which the cell is functioning properly.[33] By contrast, a cell *doesn't* have to have the particular DNA it in fact has to ensure continued proper functioning – Gibson *et al.* (2010) recently succeeded in replacing a bacterium's entire genome with one created in the laboratory, after which the cell remained capable of continuous self-replication. Particular tRNA molecules alone do not therefore necessitate the production of a certain kind of protein, even relative to those possible worlds in which the cell is functioning properly, since in some of these worlds the cell contains different DNA and therefore produces a different protein.

I don't want to suggest that every genetic explanation of phenotypic variation in contemporary biology can be explained in this way. And there may be pragmatic, or indeed *socio-political*, reasons for evaluating biological causal claims relative to larger sets of possible worlds, relative to which biological traits are caused only by large pluralities of genetic and environmental factors working together.[34] But the analysis

[33] This is plausibly what Sterelny and Kitcher have in mind when they describe genes as sufficient for many biological phenomena "relative to any *standard* environment" (Sterelny and Kitcher 1988: 349; my emphasis).

[34] Oyama reports having 'liberal colleagues' who claim "that their political views *require* them to reject biological approaches"; one of the goals of Developmental Systems Theory is apparently to

above suggests that, at least in some cases, biologists focus on DNA not out of some institutional gene-centric bias,[35] but out of a general interest in causal explanations relativized to teleological ordering sources; and Developmental Systems Theorists do not succeed, in my view, in establishing that there is anything theoretically problematic about this.

This section has touched a number of debates in diverse disciplines, from experimental philosophy to legal theory to philosophy of biology. My claim is that the phenomena described in these debates are all fundamentally effects of the same basic mechanism; one which can be effectively modelled by applying Kratzer's framework to the idea that causes necessitate their effects.

6. Causal Disagreement as Normative Disagreement

The view I have defended can be further illustrated by considering how it models the semantics of causal disagreements. Consider first the semantics of gradable adjectives – words like 'tall', 'flat', 'rich', and so on. These linguistic items are generally thought to express relations between objects and *comparison sets* (e.g. Kennedy and McNally 2005). For example, an utterance of 'Kim is tall' expresses the proposition that Kim is tall *for X*, where *X* is a set of individuals determined by the context. This means that there are two different ways of disagreeing about the truth of the proposition expressed by 'Kim is tall' in a particular context. Suppose we're in a context where 'Kim is tall' expresses the proposition that Kim is tall for a basketball player. I think 'Kim is tall' is true in such a context, whereas you think it's false. This might be because we disagree about Kim's height. But it might also be because we disagree about the average height of basketball players, and hence about how tall Kim has to *be* in order to be tall for a basketball player.

combat this perception by "question[ing] the traditional meanings of the biological to which [they] respond" (Oyama 1985: 230-1).

[35] Gannett (1999: 359), for example, tries to explain away the focus on genes as a by-product of the fact that there are better financial returns to be made on investment in genetic research.

We get exactly the same phenomenon with causal claims. Consider the following case:

> *Bankruptcy*
>
> A company is made up of low-paid workers and well-paid managers. The managers decide that they deserve more money for the work that they do and give themselves a large pay-rise; at the same time, the workers organise an unannounced strike over pay and conditions. Neither group knew of the other's plans. The company subsequently falls into financial difficulties and declares bankruptcy.

Consider a disagreement between a worker and a manager. Let's suppose that they agree on all the counterfactual facts – they agree, in particular, that the company wouldn't have had to declare bankruptcy if the strike had occurred without the pay-rise or if the pay-rise had occurred without the strike. Nevertheless, the manager makes utterances like "It wasn't the pay-rise that caused the bankruptcy, it was the strike" and the worker replies with utterances like "It wasn't the strike that caused the bankruptcy, it was the pay-rise". The source of this disagreement seems to be a *normative* disagreement – the manager believes that the pay-rise was morally required, given the work she does, and that the strike was unnecessary (the workers could have resolved their grievances in other ways); whereas the worker believes that the strike was morally required, given their working conditions and persistent managerial intransigence, and that the pay-rise was unnecessary (the managers could have done without the extra money).

These kinds of causal disagreements are pervasive in political, historical, legal and everyday discourse. But again, they're difficult to explain on standard philosophical theories of causation. How could the manager and the worker agree on all the counterfactual facts and yet fail to agree on the causal facts? The most common diagnosis is that these disagreements aren't really disagreements about *causation* at all; rather the worker and the manager really agree on all the causal facts, but disagree about something

else (such as who was *morally responsible* for the bankruptcy), and mistakenly phrase their disagreement in causal terms.[36]

On my view, we can and should take this disagreement at face value. The manager and the worker *do* disagree on the causal facts; and they do so precisely *because* they disagree on the moral facts. Although the manager and the worker agree that they're in a context where causal claims are being evaluated relative to a moral ordering source, *they disagree about what the moral ordering source is*. Hence, they disagree about which sets of possible worlds their causal claims are quantifying over. Relative to a suitable modal base, the worker thinks that every top-ranked world on the moral ordering source is one in which the strike occurs, but not every top-ranked world is one in which the pay-rise occurs. Hence she believes that, given that the pay-rise occurred, the bankruptcy *had* to occur, although the same is not true of the strike. On the other hand, the manager thinks that every top-ranked world on the moral ordering source is one in which the pay-rise occurs, but not every top-ranked world is one in which the strike occurs. Hence she believes that, given that the strike occurred, the bankruptcy *had* to occur, although the same is not true of the pay-rise.[37]

---

[36] According to Lewis, for example, the worker and the manager "disagree only about which part of the causal history is most salient for the purposes of some particular inquiry" (Lewis 1986b: 215).

[37] Does this view imply that the causal facts would have been different if the moral facts had been different (and isn't that absurd)? No, at least on a natural reading of this counterfactual. Compare with gradable adjectives again. Suppose I'm in a context where 'Kim is tall' expresses the *false* proposition that Kim is tall for a person. Now consider the sentence 'Had everyone else been shorter, Kim would have been tall'. There's a true *de dicto* reading of this counterfactual: In the relevant counterfactual worlds, Kim is tall compared to other people *in those worlds*. But the natural reading is the false *de re* reading: In the relevant worlds, Kim is tall compared to *actual* people (this latter reading is false, since Kim's *height* wouldn't have been any different if everyone else had been shorter). In this case, then, it's natural to interpret the standard of comparison as outside the scope of the modal operator. The same is true for causal claims. Suppose we're in a context in which 'The pay-rise caused the bankruptcy' expresses the proposition that the pay-rise caused the bankruptcy morally speaking, and suppose that this proposition is false. Then the natural reading of 'If the moral facts had been different, the pay-rise would have caused the bankruptcy' is the false *de re* reading: In the relevant counterfactual worlds, the pay-rise caused the bankruptcy relative to the *actual* moral ordering source. This reading is false, since the relevant modal facts wouldn't have been any different if the moral facts had been different.

To repeat, this kind of disagreement arises only in contexts where a particular ordering source is salient. The manager and the worker would presumably agree that, *physically* speaking (i.e. relative to an empty ordering source), both the pay-rise and the strike were causes of the bankruptcy. But that's not what's at issue. What's at issue is what caused the bankruptcy *morally* speaking. And they disagree about that, precisely because they disagree about the moral facts.

7. Alternatives

I've argued that we can explain a number of features of causal discourse by applying Kratzer's semantics of modality to (NC). Many alternative explanations of these phenomena have been offered in the literature. I don't have space to consider them all, but it's worth comparing my view with three popular alternative proposals – the *pragmatic* response, the *ambiguity* response and the *contrastivist* response.

Recall that subjects presented with *Collision* tend to *agree* with (3) and *disagree* with (4):

(3) Rachel's driving into the intersection caused the collision.

(4) Greta's driving into the intersection caused the collision.

On my view, this is simply because, in the relevant context, (3) is true and (4) is false. But the proponent of the *pragmatic* response disagrees. In fact, she claims, (3) and (4) are both true (in every context). But in certain contexts, an assertion of (4) would be *pragmatically inappropriate*; and that's why subjects are reluctant to endorse it. "There are ever so many reasons why it may be inappropriate to say something true", as Lewis (2000: 196) reminds us: "It might be irrelevant to the conversation, it might convey a false hint, it might be known already to all concerned", and so on.

As others have pointed out, however (see McGrath 2005; Schaffer 2013), the pragmatic response doesn't really explain the data. We aren't merely unwilling to assert (4); we're

also more than happy to assert its *negation*. The following seems like a perfectly natural thing to say in the context:

> (5) Greta's driving into the intersection didn't cause the collision. It was Rachel's driving into the intersection which caused it; Greta's driving into the intersection was just a background condition.

It doesn't follow from the fact that it's pragmatically inappropriate to utter a true sentence that it *would* be appropriate to utter its negation. If you ask me for directions and I respond with 'The earth is round', my utterance is pragmatically inappropriate, since the information I supply is irrelevant. But that doesn't mean that it *would* have been acceptable for me to respond instead with 'The earth isn't round'. So simply pointing out that it would be pragmatically inappropriate to utter (4) in some contexts doesn't explain why it *is* appropriate in those contexts to utter (5).

There are admittedly some instances of so-called *metalinguistic negation* in which the word 'not' is used to register dissatisfaction with the pragmatic propriety of a would-be utterance rather than with its literal falsity. Consider the following sentence:

> (6) The water isn't *hot*; it's scalding!

(6) is contradictory if the (contracted) 'not' is interpreted as expressing logical negation – if something is scalding, then necessarily it's hot. But there is nevertheless a consistent reading of (6), one where the 'not' is interpreted metalinguistically. On this reading, (6) says that an utterance of 'It's hot' was (or would have been) inappropriate, because it (would have) pragmatically implicated the false proposition that the water isn't scalding. One might therefore be tempted to say the same about (5). It has a true reading, one where the 'not' is interpreted metalinguistically. Perhaps, for example, (5) says that an utterance of 'Greta's driving into the intersection caused the collision' would have been inappropriate, because it would have pragmatically implicated the false proposition that Greta was *to blame* for the collision.

The only problem with this proposal is that there isn't any actual evidence that 'not' is being used metalinguistically in (5). Metalinguistic readings are fickle (see Horn 1989). Firstly, they rely on specific intonation contours. Placing the stress in a different place in (6) makes the metalinguistic reading much harder to access:

(7) # The water *isn't* hot. It's scalding.

Secondly, metalinguistic readings disappear in 'concessive' constructions like (8):

(8) # The water isn't hot, although it is scalding.

Although (6) has a clear true reading, (7) and (8) strike us as odd or even contradictory. But we don't observe the same patterns for (5):

(9) Greta's driving into the intersection *didn't* cause the collision. It was just a background condition.

(10) Greta's driving into the intersection didn't cause the collision, although it was a background condition.

Speaking for myself, I find (9) and (10) to be just as acceptable as (5). This suggests that the 'not' in (5) is not being interpreted metalinguistically. These tests are not decisive, of course, but they work well enough to make us suspicious of unqualified appeals to pragmatics.

The second response, the *ambiguity* response, postulates an ambiguity in the word 'cause'. This view "treats our ordinary talk of causation as shifting between two distinct 'concepts' of cause", resulting in "an ambiguity that could be resolved if each concept was given a different name" (Godfrey-Smith 2009: 327). One is the "broad and non-discriminatory notion of causation" (Lewis 1986a: 162) and the other is a narrower "selective notion of causation" (Broadbent 2012: 463). We might call the former 'causal influence' and the latter 'explanatory relevance', for example (Strevens 2008). Either way, the idea is that Greta's driving into the intersection 'caused' the collision in the

broad but not the narrow sense, whereas Rachel's driving into the intersection 'caused' the collision in both the broad and the narrow sense. (5) is an acceptable thing to say because it has a true interpretation, one where 'cause' is interpreted narrowly.

Again, the problem with this view is that there is no actual *evidence* that 'cause' is ambiguous in this way. An ambiguity between a broad and a narrow understanding of a word is known as a *privative* ambiguity in linguistics. Zwicky and Sadock (1975) suggest a number of tests for determining whether a term is privatively ambiguous. One of these is to check whether it is "possible, without contradiction, to assert the general while denying the specific" (Zwicky and Sadock 1975: 7). Compare the following two sentences:

(11) The aeroplane climbed, but it didn't *climb*.

(12) # The aeroplane crashed, but it didn't *crash*.

There is no way of understanding (12) on which it isn't a contradiction. But there is a way of understanding (11) on which it isn't a contradiction. This suggests that 'climb' is ambiguous between a narrow literal meaning (where climbing requires arms and legs) and a broader, more metaphorical meaning.

Now consider the following sentence:

(13) Greta's driving into the intersection caused the collision, but it didn't *cause* the collision.

Speaking for myself, I find it hard to understand (13) as anything other than a straightforward contradiction. This suggests that 'cause' is not ambiguous. Again, this test is not decisive, but it works well enough to make us suspicious of unqualified appeals to ambiguity.

The last response I'll consider is the *contrastivist* response (Schaffer 2005, 2013; Hitchcock 2007). According to contrastivism, causation is a four-place, *contrastive* relation – roughly speaking, *C* rather than *C** caused *E* rather than *E** if and only if, had

*C** occurred instead of *C*, *E** would have occurred instead of *E*.[38] Sentences of the form '*C* caused *E*' express contrastive propositions, on this view, where the contrast events are determined by the conversational context. The idea is that "'deviant' events tend to leap out as especially salient to people and tend to trigger thoughts of the more normal alternative, while 'default' events tend to duck out of view and not trigger thoughts about alternatives at all" (Blanchard and Schaffer forthcoming). For example, Rachel's driving into the intersection is a 'deviant' event, since she had a red light. It therefore triggers thoughts of a more 'normal' alternative, namely Rachel's stopping at the intersection. Hence an utterance of (3) is naturally interpreted as expressing the (true) proposition that Rachel's driving into the intersection rather than stopping caused there to be a collision rather than no collision. But Greta's driving into the intersection was a 'default' event, since she had a green light. Hence it doesn't trigger thoughts of any alternatives. It follows that an utterance of (4) cannot be interpreted as expressing a contrastive proposition, "nor is it obvious what if any interpretation it should receive" (Schaffer 2013: 52).

I think the contrastivist is wrong about the data here. If there is indeed no salient alternative to Greta's driving into the intersection, (4) is *uninterpretable* on the contrastive account, since one of the arguments of the function expressed by 'cause' remains unspecified by the context. But if (4) is uninterpretable then so is its negation, for the same reason. Hence (5) is also uninterpretable according to the contrastivist – it's neither true nor false, since it fails to express a proposition at all. This seems like the wrong result. Speaking for myself, (5) just seems straightforwardly *true*. Schaffer himself seems to acknowledge as much in a footnote (he is concerned with a different sentence, but the same considerations apply):

[38] 'Roughly speaking', because this simple counterfactual analysis runs into problems in familiar cases involving redundant causation. These problems won't be relevant in what follows.

> *Lacuna*: if [4] does not receive any natural interpretation than [sic] its denial should not either, which does not quite fit that data…So it would be smoother for me to say that [4] does receive some interpretation as a contrastive falsehood…But I do not currently have any contrastive falsehood to suggest for the role (Schaffer 2013: 61-2).

These considerations aren't decisive, but again, they should make us suspicious of unqualified appeals to contrastivism to explain these phenomena.[39]

I've considered three alternative explanations of our judgements concerning (3), (4) and (5). Though I don't purport to have provided any knock-down arguments, I've shown that there are serious problems with each of them. This, at the very least, is reason to take the view I have defended seriously.

8. Conclusion

Mackie famously argued that "causal statements are commonly made in some context, against a background which includes the assumption of some *causal field*" (Mackie 1974: 34). The 'causal field', for Mackie, is not "part of a cause, but is rather a background against which the causing goes on" (Mackie 1974: 63). The view defended in this paper can be seen as a contemporary development of Mackie's insight, obtained by combining the ancient idea that causes necessitate their effects with a familiar framework borrowed from Kratzer's theory of modality.

This paper has largely focussed on the semantics of causal language. I take my proposal to be consistent with a number of different metaphysical theories of causation. In particular, I take my proposal to be neutral on the question of whether the 'necessary connections' between causes and their effects should be grounded in patterns of occurrences or *vice versa* (see Beebee 2000, 2006), as well as the question of whether (NC) can be turned into a reductive analysis of causation by the addition of further

[39] For criticisms of other applications of the contrastivist framework, see Chapter Three, Section 5, below.

necessary conditions. One might also accept that causal claims often *are* evaluated relative to bouletic, deontic and teleological ordering sources in everyday discourse, but nevertheless insist that they *shouldn't* be evaluated in this way, at least in specialist domains like science and the law. The question of whether such stipulated constraints on the context-sensitivity of causal language have their practical merits is beyond the scope of semantics, and therefore beyond the scope of this paper. It's worth pointing out that biology and the law have arguably managed just fine without them, however. Indeed it's worth remembering why natural languages contain context-sensitive elements in the first place: They allow us to exploit rich and subtle features of the circumstances in which we find ourselves to successfully communicate a large variety of different kinds of content with minimal linguistic resources. What's not to like that about that?

References

Anscombe, G. E. M. (1971). *Causality and determinism*. Cambridge: Cambridge University Press.

Beebee, H. (2000). The non-governing conception of laws of nature. *Philosophy and Phenomenological Research 61*(3), 571-594.

Beebee, H. (2006). Does anything hold the universe together?. *Synthese 149*(3), 509-533.

Blanchard, T. and Schaffer, J. (forthcoming). Cause without default. In H. Beebee, C. Hitchcock and H. Price (eds.), *Making a difference*, Oxford: Oxford University Press.

Broadbent, A. (2012). Causes of causes. *Philosophical Studies 158*(3), 457-476.

Cappelen, H. and Lepore, E. (2005). *Insensitive semantics: A defense of semantic minimalism and speech act pluralism*. Oxford: Blackwell.

Clarke, R., Shepherd, J., Stigall, J., Waller, R. R. and Zarpentine, C. (2015). Causation, norms, and omissions: A study of causal judgments. *Philosophical Psychology 28*(2), 279-293.

Davidson, D. (1980). Mental events. In D. Davidson, *Essays on actions and events*, Oxford: Oxford University Press.

Edgarton, H. W. (1924). Legal cause. *University of Pennsylvania Law Review 72*(3-4), 211-244, 343-375.

Faye, J. (2014). *The nature of scientific thinking*. London: Palgrave Macmillan.

Fischer, D. A. (2006). Insufficient causes. *University of Kentucky Law Review 94*, 277-317.

Fumerton, R. and Kress K. (2001). Causation and the law: Preemption, lawful sufficiency, and causal sufficiency. *Law and Contemporary Problems 64*(4), 83-105.

Gannett, L. (1999). What's in a cause?: The pragmatic dimensions of genetic explanations. *Biology and Philosophy 14*(3), 349-373.

Gibson, D. *et al.* (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science 329*(5987), 52-56.

Godfrey-Smith, P. (2009). Causal pluralism. In H. Beebee, P. Menzies and C. Hitchcock (eds.), *The Oxford handbook of causation*, Oxford: Oxford University Press.

Green, L. (1927). *Rationale of proximate cause*. Kansas City, MO: Vernon Law Book Company.

Green, L. (1929). Are there dependable rules of causation?. *University of Pennsylvania Law Review 77*(5), 601-628.

Hacquard, V. (2009). On the interaction of aspect and modal auxiliaries. *Linguistics and Philosophy 32*(3), 279-312.

Hart, H. L. A. and Honoré, T. (1985). *Causation in the law (2nd ed.)*. Oxford: Clarendon Press.

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *The Philosophical Review 116*(4), 495-532.

Hitchcock, C. and Knobe, J. (2009). Cause and norm. *The Journal of Philosophy 106*(11), 587-612.

Hobbes, T. ([1655]1839). Elements of philosophy concerning body. In W. Molesworth (ed.), *The English works of Thomas Hobbes of Malmesbury*, London: Bohn.

Hoffmann, L. (2011). Causation. In R. Goldberg (ed.), *Perspectives on causation*, Oxford: Hart Publishing.

Horgan, T. and Timmons, M. (2006). Expressivism, yes! Relativism, no!. In R. Shafer-Landau (ed.), *Oxford studies in metaethics (vol. 1)*, Oxford: Clarendon Press.

Horn, L. (1989). *A natural history of negation*. Chicago: University of Chicago Press.

Hume, D. ([1738]1978). *A treatise of human nature (ed. P. H. Nidditch)*. Oxford: Clarendon Press.

Kennedy, C. and McNally, L. (2005). Scale structure and the semantic typology of gradable predicates. *Language 81*(2), 345-381.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences 33*(4), 315-365.

Kratzer, A. (1981). The notional category of modality. In H. J. Eikmeyer and H. Rieser (eds.), *Words, worlds and contexts*, New York: de Gruyter.

Kratzer, A. (2012). *Modals and conditionals*. Oxford: Oxford University Press.

Krifka, M., Pelletier, F. J., Carlson, G. N., ter Meulen, A., Chierchia, G. and Link, G. (1995). Generiticity: An introduction. In G. N. Carlson and F. J. Pelletier (eds.), *The generic book*, Chicago: University of Chicago Press.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1986a). Causation. In D. Lewis, *Philosophical papers (vol. 2)*, Oxford: Oxford University Press.

Lewis, D. (1986b). Causal explanation. In D. Lewis, *Philosophical papers (vol. 2)*, Oxford: Oxford University Press.

Lewis, D. (1986c). *On the plurality of worlds*. Oxford: Blackwell.

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy 97*(4), 182-197.

Loewer, B. (2007). Mental causation, or something near enough. In B. P. McLaughlin and J. D. Cohen (eds.), *Contemporary debates in philosophy of mind*, Oxford: Blackwell.

Lunney, M. and Oliphant, K. (2013). *Tort law: Text and materials (5th ed.)*. Oxford: Oxford University Press.

MacFarlane, J. (2005). Making sense of relative truth. *Proceedings of the Aristotelian Society 105*(1), 321-339.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly 2*(4), 245-264.

Mackie, J. L. (1974). *The cement of the universe*. Oxford: Clarendon Press.

Malebranche, N. ([1674]1997). *The search after truth (trans. T. M. Lennon and P. J. Olscamp)*. Cambridge: Cambridge University Press.

McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies 123*(1-2), 125-148.

McKay, T. (2006). *Plural predication*. Oxford: Oxford University Press.

Mellor, D. H. (1995). *The facts of causation*. London: Routledge.

Mill, J. S. ([1843]1868). *A system of logic, ratiocinative and inductive (vol. 1)*. London: Longmans, Green, Reader, and Dyer.

Mumford, S. and Anjum, R. L. (2011). *Getting causes from powers*. Oxford: Oxford University Press.

Oyama, S. (1985). *The ontogeny of information: Developmental systems and evolution*. Durham, NC: Duke University Press.

Oyama, S. (2000). Causal democracy and causal contributions in developmental systems theory. *Philosophy of Science 67*(3), S332-S347.

Oyama, S., Griffiths, P. E. and Gray, R. D. (2001). *Cycles of contingency: Developmental systems and evolution*. Cambridge, MA: The MIT Press.

Plato (2002). *Phaedo (trans. D. Bostock)*. Oxford: Clarendon Press.

Reid, T. ([1788]2010). *Essays on the active powers of man (ed. K. Haakonssen and J. A. Harris)*. Pennsylvania, PA: The Pennsylvania State University Press.

Schaffer, J. (2005). Contrastive causation. *The Philosophical Review 114*(3), 327-358.

Schaffer, J. (2013). Causal contextualism. In M. Blaauw (ed.), *Contrastivism in philosophy*, New York: Routledge.

Stalnaker, R. C. (1981). A defense of conditional excluded middle. In W. Harper, R. C. Stalnaker and G. Pearce (eds.), *Ifs*, Dordrectht: D. Reidel.

Stapleton, J. (2001). Legal cause: Cause-in-fact and the scope of liability for consequences. *Vanderbilt Law Review 54*(3), 941-1024.

Stapleton, J. (2008). Choosing what we mean by 'causation' in the law. *Missouri Law Review 73*(3), 433-480.

Stapleton, J. (2010). Factual causation. *Federal Law Review 38*(1), 467-484.

Stegmann, U. E. (2012). Varieties of parity. *Biology and Philosophy 27*(6), 903-918.

Stein, N. (forthcoming). Causes and categories. *Noûs*.

Sterelny, K. and Kitcher, P. (1988). The return of the gene. *The Journal of Philosophy 85*(7), 339-361.

Strevens, M. (2008). *Depth: An account of scientific explanation*. Cambridge, MA: Harvard University Press.

Szabó, Z. G. and Knobe, J. (2013). Modals with a taste of the deontic. *Semantics and Pragmatics 6*(1), 1-42.

Unger, P. (1977). The uniqueness in causation. *American Philosophical Quarterly 14*(3), 177-188.

Waters, C. K. (2007). Causes that make a difference. *The Journal of Philosophy 104*(11), 551-579.

Weber, M. (forthcoming). Causal selection versus causal parity in biology: Relevant counterfactuals and biologically normal interventions. In C. K. Waters, M. Travisano and J. Woodward (eds.), *Philosophical perspectives on causal reasoning in biology*, Minneapolis: University of Minnesota Press.

Wright, R. W. (1985). Causation in tort law. *California Law Review 73*(6), 1735-1828.

Wright, R. W. (2001). Once more into the bramble bush: Duty, causal contribution, and the extent of legal responsibility. *Vanderbilt Law Review 54*(3), 1071-1132.

Wright, R. W. (2007). Acts and omissions as positive and negative causes. In J. W. Neyers, E. Chamberlain and S. G. A. Pitel (eds.), *Emerging issues in tort law*, Oxford: Hart Publishing.

Wright, R. W. (2011). The NESS account of natural causation: A response to criticisms. In R. Goldberg (ed.), *Perspectives on causation*, Oxford: Hart Publishing.

Zwicky, A. and Sadock, J. (1975). Ambiguity tests and how to fail them. In J. Kimball (ed.), *Syntax and semantics (vol. 4)*, New York: Academic Press.

Chapter Two

Alternative Possibilities in Context

Abstract: This paper defends a contextualist dissolution of the freedom and determinism problem that combines the ancient idea that free action requires the possibility of acting otherwise with Angelika Kratzer's semantics of modality. On the resulting view, the truth of everyday ascriptions of freedom is perfectly compatible with the truth of determinism. This contextualist proposal has other virtues too – it can explain what's wrong with van Inwagen's 'consequence argument', it can explain what's wrong with the so-called 'conditional analysis' of the principle of alternative possibilities, it can explain why ascriptions of freedom are sensitive to normative facts, and it can clarify the relationship between freedom and moral responsibility.

1. Introduction

Many human actions are free. I acted freely when I made coffee this morning, when I answered my phone, when I asked a friend for the time. To doubt that would be absurd. At any rate, to doubt it in any serious and lasting way would be absurd. It is a Moorean fact that some human actions are free. It is one of those things that we know better than we know the premises of any philosophical argument to the contrary.[1]

And yet no sooner do we reflect on our place in the physical world than we meet a compelling argument that no action is free. Centuries of scientific investigation have lent greater and greater credence to *physical determinism*, the hypothesis that every concrete

[1] This is a deliberate paraphrase of the opening paragraph of Lewis (1996). As will become clearer below, there are a number of interesting parallels between the view defended here and Lewis's solution to the problem of scepticism.

event is metaphysically necessitated by the fundamental physical laws and the fundamental physical state of the universe at a time.[2, 3] If determinism is true, every action anyone has ever performed was metaphysically necessitated by the laws of nature and the state of the universe immediately after the big bang. Since we have no control over either of these things, it seems to follow that no-one could have acted otherwise than they eventually did, from which it seems to follow that no action is free. My perception of myself as a free agent is but an illusion – indeed, that very perception itself was determined long before I was even born!

The argument I have just sketched has three premises:

> DETERMINISM (D): The laws of nature are deterministic.
>
> CONSTRAINT (C): If the laws of nature are deterministic, no-one could have acted otherwise than they actually did.
>
> ALTERNATIVE POSSIBILITIES (AP): An agent acts freely only if she could have acted otherwise.

The *freedom and determinism problem*, as I shall refer to it, is simply this: (D), (C) and (AP) are jointly inconsistent with (F):

> FREEDOM (F): Some actions are free.

---

[2] Or, more precisely, the fundamental state of any maximal space-like hypersurface.

[3] The popular wisdom is that physical theories were fully deterministic until quantum mechanics introduced the possibility of genuinely indeterministic laws of nature. But in fact the exact opposite is closer to the truth – determinism is uncontroversially false in classical mechanics and, only slightly more controversially, in general relativity too (see Earman 1986). The laws of orthodox quantum mechanics, on the other hand, are perfectly deterministic (see Wallace ms). Some theorists have responded to perceived conceptual difficulties with quantum mechanics by postulating an *additional* indeterministic dynamical process, called 'collapse'; but there are other deterministic alternatives (e.g. De-Broglie-Bohm theory) as well as interpretations of quantum mechanics that take it seriously on its own terms (e.g. the Everett or 'many worlds' interpretation). Ironically, then, quantum mechanics is arguably "one of the best prospects for a genuinely deterministic theory in modern times" (Hoefer 2016).

A solution to the freedom and determinism problem must therefore explain which of these premises ought to be abandoned. *Hard incompatibilists* reject (F), and simply accept that no action is free.[4] *Libertarians* reject (D), and argue that free actions are not nomologically determined by prior states of affairs.[5] *Compatibilists* come in two varieties. Some reject (C), and insist that it doesn't follow from the fact that an action was nomologically determined by prior states of affairs that the agent couldn't have acted otherwise.[6] Others deny (AP) – according to a tradition going back to Hobbes, for example, an agent acts freely so long as he *does what he wills*, regardless of whether he could have acted otherwise.[7]

Some philosophers – call them *deflationists* – don't straightforwardly fit into any of these categories. Deflationists deny that there is a freedom and determinism problem to be solved at all. Any appearance to the contrary is down to a failure to pay close enough attention to the semantic variability of the word 'free'. There is one sense of 'free' according to which an agent acts freely only if determinism is false; and there is a *different* sense of 'free' according to which many actions are free, regardless of whether determinism is true. The argument above seems to work only by equivocating between these two interpretations. According to Hume, for example, once we distinguish "betwixt the liberty of spontaneity and…the liberty of indifference" (Hume 1739: 407), it's clear "that all men have ever agreed in the doctrine both of necessity and of liberty, according

[4] E.g. d'Holbach (1770); Pereboom (2001).

[5] See especially Balaguer (2002); Clarke (2003); Kane (1985, 1996); Wiggins (1973).

[6] Lewis (1981) can be interpreted as defending a view of this kind; although see section 3, below. See also Fara (2008) and Vihvelin (2004).

[7] "[I]t cannot be conceived that there is any Liberty greater, than for a Man to do what he will" (Hobbes 1654: 479). Frankfurt (1971) points out that a psychotic who acts on a desire to jump from a window, or an addict who acts on a desire to take heroin, or a kleptomaniac who acts on a desire to steal a packet of crisps, can properly be described as 'doing what he wills', even though we would not normally regard him as acting freely. To account for this, Frankfurt appeals to *higher-order* desires – a free action is one caused by a desire that the agent desires to be effective. But the basic strategy here is the same – to analyse freedom in terms of the 'actual sequence' of events in the causal history of the action, rather than in terms of the agent's ability to do otherwise (see Fischer and Ravizza 1998; see also section 6, below).

to any reasonable sense, which can be put on these terms; and that the whole controversy has hitherto turned merely upon words" (Hume 1748: 81).

Of course, metaphysical theorising would be an easy task indeed if every problem could be dissolved in this way. It might be tempting to explain away the intractability of a metaphysical dispute by postulating the existence of a semantic mistake, but then philosophy is hard – one should have some independent evidence for thinking that everyone is simply talking past one another. That said, I believe that a kind of deflationist dissolution of the freedom and determinism problem can be defended. Van Inwagen (1983: 8) once said that "[f]ree will…is to be defined in terms of 'can'".[8] But despite this, there have been few attempts to apply the orthodox semantics of 'can' to the freedom and determinism problem. This is what I propose to do in this paper. In section 2, I defend a contextualist semantics of ascriptions of freedom to actions by applying Kratzer's semantics of modality to (AP). Sections 3 and 4 use this semantics to expose flaws in popular arguments on both sides of the compatibilism/incompatibilism divide, and section 5 uses it to explain the sensitivity of our judgements about freedom to *normative* features of cases. Finally, in sections 6 and 7, I explore the novel consequences of this contextualist proposal for how to think about the relationship between freedom and moral responsibility, using as an illustrative case study the long-running debate in legal theory over the proper interpretation of the common law defence of duress.

2. Freedom in Context

What is the logical form of the sentence 'I could have acted otherwise'? The surface structure would suggest a simple subject-predicate form, constructed out of a noun-phrase 'I' and a complex verb-phrase 'could have acted otherwise'. On the orthodox approach to the semantics of modal auxiliaries (Kratzer 1977, 1981), however, the 'could' in 'I could have acted otherwise' is not part of a complex verb phrase 'could have acted

[8] He now prefers the locution 'is able to' (van Inwagen 2008) – although see section 4, below.

otherwise', but is rather a sentential operator ranging over the whole sentence 'I acted otherwise'.[9] This operator expresses a function with three arguments: A proposition (e.g. the proposition that I acted otherwise) and two sets of propositions, the modal base **B** and the ordering source **O**. Intuitively, one can think of the modal base as the set of things we're 'holding fixed' in the context, and the ordering source as a set of salient 'ideal' states of affairs, where 'ideal' here could mean morally or socially or legally ideal, or 'ideal' in the sense of maximally normal or stereotypical, or ideal in light of a set of goals or desires or teleological ends. These two parameters pick out a set of possible worlds – the *accessible* worlds – in two stages. First, we take $\cap\mathbf{B}$, the set of possible words in which all the propositions in **B** are true. Next, we use **O** to define a preorder $<_{\mathbf{O}}$ over $\cap\mathbf{B}$, as follows: $w <_{\mathbf{O}} u$ if and only if $\{p : p \in \mathbf{O} \,\&\, w \in p\} \subset \{p : p \in \mathbf{O} \,\&\, u \in p\}$. Relative to a context, modal claims quantify over max(**B**, **O**), the set of maximal elements of this ordering (where $w \in \max(\mathbf{B}, \mathbf{O})$ if and only if $w \in \cap\mathbf{B}$ and there is no $u \in \cap\mathbf{B}$ such that $w <_{\mathbf{O}} u$). Hence, I could have acted otherwise relative to **B** and **O** if and only if I did act otherwise in at least one element of max(**B**, **O**).

According to Kratzer's theory, then, the right-hand-side of (AP) is context-sensitive – the sentence 'She could have acted otherwise' expresses different propositions on different occasions of use, depending on the modal base and ordering source provided by the context. I think the same is true of the left-hand-side of (AP). On the view I will defend, an action isn't simply free *simpliciter*, but only *relative* to a modal base and ordering source. Properly spelled out, therefore, (AP) should read as follows:

> (AP): S's action was free relative to **B** and **O** only if S acts otherwise in some element of max(**B**, **O**).[10]

[9] Although see Vetter (2015) for a dissenting view.

[10] Note that (AP) is just a necessary condition on free action – it's perfectly consistent with (AP) that an agent acts freely only if she is the 'ultimate originator' of her actions, for example (see Kane 1996).

We're now in a position to see, at least in broad outline, how the freedom and determinism problem can be dissolved. Suppose I drank coffee this morning. Was my action 'free'? Well, that depends on the context. When we think about this action from the perspective of fundamental physics, we might find ourselves in a context relative to which the laws of nature and the initial state of the universe are held fixed – that is, included in the modal base. If determinism is true, there is only one possible world in which the laws of nature and initial state of the universe are the same as they actually are, namely the actual world. So relative to this *demanding* context, I couldn't have acted otherwise than I actually did, and hence by (AP), my drinking coffee this morning wasn't free. But many everyday ascriptions of freedom are made in much more *permissive* contexts, relative to which much less is held fixed. For example, one might be interested in whether my making coffee this morning was free, holding fixed only certain facts about my psychological dispositions (am I addicted to coffee?), facts about the external conditions (was there anything else to drink?), or some combination of the two. Relative to *these* contexts, many more worlds are accessible, and so it doesn't necessarily follow from (AP) that my drinking coffee this morning wasn't free. In this way, we can preserve the intuition that many everyday ascriptions of freedom are *true*, while at the same time insisting that to utter 'My action was free' and 'I couldn't have acted otherwise', *in the same context*, is to contradict oneself. (F) is true in some contexts and (C) is true in others; the lesson of the freedom and determinism problem is just that there are no contexts relative to which *both* are true.

The basic idea here is not particularly new. Kratzer herself hints at it in the following passage:

> Suppose a judge asks himself whether a murderer could have acted otherwise than he eventually did…Given the whole situation of the crime, which includes of course all the dispositions of the murderer, this man could not have acted otherwise than as he did…[But t]he answer to the question of the judge is *not* trivial…what the judge probably meant was:

> Given such and such aspects of the situation, could the murderer have acted otherwise than he eventually did? (Kratzer 1977: 343).

Though he didn't have the resources of possible-worlds semantics, Nowell-Smith expressed the same basic insight some years earlier:

> The fallacy in this argument lies in supposing that, when we say 'A could have acted otherwise', we mean that A, *being what he was and being placed in the circumstances in which he was placed*, could have done something other than what he did. But in fact we never do mean this (Nowell-Smith 1948: 49; my emphasis).

List (2014: 169), citing Kratzer, defends a similar view, which he claims is "broadly consistent with some established accounts of the ordinary meaning of 'can'". And Hawthorne points to similarities between the freedom and determinism problem and the problem of scepticism as evidence that a contextualist dissolution of the freedom and determinism problem "deserves to be taken seriously" (Hawthorne 2001: 77). Despite these hints, however, the consequences of such a contextualist treatment of (AP) for the broader compatibilism/incompatibilism debate remain largely underexplored. It's the goal of this paper to describe these consequences in more detail.

3. The Consequence Argument

According to the contextualist proposal sketched above, there are contexts relative to which (C) is false. Relative to these 'permissive' contexts, it's consistent with (AP) that many actions are free, even if the laws of nature are deterministic. The idea is that (C) seems so plausible only because mere mention of determinism is often sufficient to *change* the context to one relative to which it expresses a truth, namely one in which the laws of nature and the state of the universe at some time in the past are included in the modal base.

Van Inwagen (1983) has defended a much-discussed argument for (C), however, which he calls the 'consequence argument'. There are many different versions of the

consequence argument, but the simplest version goes like this. Let P be a proposition that describes the complete state of the universe at some time before my birth and let L be the conjunction of all the fundamental laws of nature. If I could have acted otherwise and determinism is true, then I could have rendered L&P false. But I couldn't have rendered L false – I'm not a miracle worker – and I couldn't have rendered P false – the past, after all, is past. So I couldn't have rendered L&P false. So if determinism is true, I couldn't have acted otherwise.

Although the consequence argument is superficially compelling, it is in fact no threat to the contextualist treatment of (AP). Explaining why this is so will take some time; but given the volume of discussion the consequence argument has generated, this will be time well spent. Predictably, a great deal turns on what exactly is meant by 'render false' here. Following Lewis (1981), let's introduce some unambiguous terminology: "Let us say that an event would falsify a proposition iff, necessarily, if that event occurs then that proposition is false" (Lewis 1981: 119). In other words, an event *E* 'falsifies' a proposition *p* if and only if there is no possible world in which *E* occurs and *p* is true.[11] Now consider the following interpretation of 'can render false':

(1) S can render *p* false if and only if S can perform an action which would falsify *p*.

If 'render false' is understood in this way, the consequence argument is invalid. Assuming determinism, there is no possible world in which L&P is true and S performs a different action to the one she actually performs. Hence if S can act otherwise, she can perform an action which would falsify L&P. I accept this conclusion. But it *doesn't* follow from this that S can perform an action which would falsify either L or P. Performing an action that is incompatible with the truth of L&P isn't the same as performing an action which is either incompatible with the truth of L or incompatible with the truth of P. Suppose I

[11] Note that every event falsifies the proposition that 2+2=5, on this definition – I'll ignore this complication in what follows.

made coffee this morning. If I had made tea instead, my making tea would have falsified L&P, assuming determinism. But it wouldn't have falsified L, because there are many possible worlds in which I make tea and L is true, ones in which the past differs from the actual past. Nor would it have falsified P, because there are many possible worlds in which I make tea and P is true, ones in which some of the actual laws of nature are false. Put another way: It's true that if I could have acted otherwise, I could have performed an action such that had I performed it, either L or P would (have to)[12] have been *false*; but it *doesn't* follow from this that I could have performed an action such that, had I performed it, either L or P would (have to) have been *falsified*, whether by my action or by any other event.

It's just as well, then, that van Inwagen does not subscribe to (1) as a definition of 'can render false'. Instead, he proposes the following definition:

> [W]e may define '*s* can render *p* false' as follows: It is within *s*'s power to arrange or modify the concrete objects that constitute his environment in some such way such that it is not possible in the broadly logical sense that he arrange or modify those object in that way *and the past have been exactly as it in fact was* and *p* be true. (Van Inwagen 1983: 68; my emphasis).

In other words:

> (2) S can render *p* false if and only if S can perform an action *A* such that *A* and the actual state of the universe at all times in the past would (jointly) falsify *p*.

As van Inwagen correctly argues, it follows from (2) that if S could have acted otherwise, she could have rendered L false. Suppose I made coffee this morning and determinism is true; then in every possible world in which I don't make coffee and the past is exactly the same as it actually is, L is false; and so if I could have acted otherwise, I could have

[12] On why this parenthetical is required to bring out the intended readings of so-called 'backtracking counterfactuals', see Arregui (ms).

performed an action which, together with the actual state of the universe at all times in the past, would have falsified L.

Van Inwagen thinks that it is obvious that I can't render L false. But this is because he is thinking of 'render false' in terms of (1) rather than his own definition, (2). Consider the following passage, for example:

> Suppose a bureaucrat of the future orders an engineer to build a spaceship capable of travelling faster than light. The engineer tells the bureaucrat that it's a law of nature that nothing travels faster than light. The bureaucrat concedes this difficulty, but counsels perseverance: "I'm sure", he says, "that if you work hard and are very clever, you'll find some way to go faster than light, even though it's a law of nature that nothing does." Clearly his demand is simply incoherent. (Van Inwagen 1983: 62).

The bureaucrat's demand is indeed incoherent. But that is because the bureaucrat is asking the engineer to perform an action which would falsify L *by itself*. He is asking, in other words, for the engineer to perform an action that doesn't occur in *any* possible world in which the laws of nature are true. All of van Inwagen's other examples (asking someone to build a device which violates conservation of momentum, for example) also take this form. But as we saw above, if I could have acted otherwise, it *doesn't* follow that I could have performed an action which *by itself* would have falsified L. For the consequence argument to succeed, then, van Inwagen needs to convince us that it is incoherent to demand that someone perform an action which *together with the actual state of the universe at all times in the past* would falsify L. And he gives us precisely no reason to suppose that this is the case. To the contrary, we make such demands of people *all the time* – if I demand that you return the book that I lent to you last week, and you refuse, then my demand (as it turns out) was a demand for you to perform an action which would, together with the actual state of the universe at all times in the past, have falsified L, assuming determinism. Compared with the bureaucrat's demand for faster-than-light travel, however, my demand that you return my book seems perfectly

reasonable. It *might* after all be incoherent – error theories are always available – but it's no use arguing for this by pointing out the absurdity of demanding faster-than-light travel.

In summary, van Inwagen's consequence argument equivocates between two different interpretations of what it is to 'render false' a proposition – on one interpretation the argument is invalid, and on the other the argument is valid but, for all van Inwagen succeeds in establishing, unsound. There are, admittedly, many more versions of the consequence argument. Some of these do not make use of the concept of 'rendering a proposition false', preferring instead to talk of the possible worlds to which an agent has 'access' (van Inwagen 1983: 78-93) or the propositions about whose truth an agent 'has a choice' (van Inwagen 1983: 93-105). I will resist the temptation to discuss these alternative versions (one could easily write an entire thesis on the consequence argument), but van Inwagen himself admits that these arguments make "essentially the same point", differing only "in structure and vocabulary" (van Inwagen 1983: 56).[13] I therefore conclude that the consequence argument fails to establish that (C) is true, and therefore poses no threat to the contextualist approach to (AP) defended here.

Before moving on, it's worth comparing my response to the consequence argument with a superficially similar, but crucially different, response due to Lewis (1981). Lewis also distinguishes between two different interpretations of 'render false':

> Let us say that I could have rendered a proposition false in the weak sense iff I was able to do something such that, if I did it, the proposition would have been falsified (though not necessarily by my act, or by any event caused by my act). And let us say that I could have rendered a proposition false in the strong sense iff I was able to do something such that, if I

[13] See McKay and Johnson (1996) for an argument against 'rule β', one of two inference rules that, according to van Inwagen, form part of the logic of the sentential operator N, where N*p* abbreviates '*p*, and no one has, or ever had, any choice about whether *p*'. McKay and Johnson point out that rule β implies another rule, *agglomeration* – N*p* & N*q* ⊨ N(*p*&*q*) – and agglomeration is invalid, for basically the same reasons that 'I can falsify L&P, therefore I can falsify L or I can falsify P' is invalid.

> did it, the proposition would have been falsified either by my act itself or by some event caused by my act. (Lewis 1981: 120).

Lewis then distinguishes between two different theses:

> The Weak Thesis, which…I accept, is the thesis that I could have rendered a law false in the weak sense. The Strong Thesis, which I reject, is the thesis that I could have rendered a law false in the strong sense. (Lewis 1981: 120).

Lewis then claims that an ability to act otherwise entails the Weak Thesis but not the Strong Thesis. This conclusion relies on Lewis's theory of counterfactuals and his account of the closeness ordering on possible worlds (Lewis 1973, 1979). On Lewis's view, a counterfactual like 'If I had raised my hand, the taxi would have pulled over' is true if and only if the taxi pulls over in every *closest* accessible world in which I raise my hand.[14] The closest accessible worlds in which I raise my hand, according to Lewis, are ones in which the state of the universe evolves exactly as it in fact did until few seconds ago, at which a point a small 'miracle' occurs in my brain, which causes me to decide to raise my arm, which I then go on to do. Thus if I had raised my arm, a 'miracle' would have occurred, but that 'miracle' wouldn't have been, nor would it have been caused by, the raising of my arm. It follows that if I could have raised my arm, I could have rendered L false in the weak but not the strong sense; which is just as well, according to Lewis, since an ability to render L false in the strong sense would be "[a] marvelous power indeed!" (Lewis 1981: 114). "[M]y ability to raise my hand", by contrast, "confers no marvelous ability to break a law, even though a law would be broken if I did it" (Lewis 1981: 116).

Lewis is right, of course, that we are not able to render the laws false in the strong sense – we are not able to perform actions which would falsify L (by themselves). But contrary

[14] I'm assuming here that, for every proposition φ, there is a non-empty set of *closest* φ-worlds, rather than an infinite sequence of closer and closer φ-worlds. Lewis thinks there are good reasons not to make this assumption, but they won't be relevant for our purposes.

to what he claims, *we are not able to render L false in the weak sense either*.[15] According to Lewis, if I am able to raise my arm, then I have the ability to perform an action such that, had I performed it, a few neurons in my brain would have fired slightly differently than they actually did. But this event of the neurons firing – call it *N* – would *not* have falsified any law of nature, had it occurred. There are plenty of possible worlds in which the laws of nature are true and *N* occurs, ones in which the past is different to the actual past. Lewis calls *N* a 'miracle', but it's *not* a miracle in the sense that faster-than-light-travel is a miracle. Rather, what Lewis means by 'miracle' here is something like this: There is no possible world in which *N* occurs, *the state of the universe immediately preceding N obtains*, and L is true. In other words, an ability to act otherwise, according to Lewis, amounts to an ability to perform an action such that, were it to be performed, an event would occur which *together* with the state of the universe immediately prior to its occurrence would falsify L. This is *not* an ability to render L false in the weak sense, as Lewis actually defines it. So both the Strong Thesis and the Weak Thesis are false.

Suppose then that we amend Lewis's definitions of the weak and strong senses of 'can render false' as follows:

> (3) S could have rendered *p* false in the weak sense if and only if S could have performed an action *A* such that, had *A* occurred, an event would have occurred which together with the state of the universe immediately prior to its occurrence would have falsified L.
>
> (4) S could have rendered *p* false in the strong sense if and only if S could have performed an action *A* such that, had *A* occurred, either *A* or an event caused by *A*, together with the state of the universe immediately prior to its occurrence, would have falsified L.

[15] See Beebee (2003) on this point.

Under (3), the Weak Thesis is perfectly true, on Lewis's account of closeness – we *can* perform actions such that, were we to perform them, an event would occur which would, together with the state of the universe immediately preceding it, falsify a law of nature. And under (4), the Strong Thesis is false, on Lewis's account of closeness. But it's certainly not a *crazy* thesis, so-understood; what it effectively says is that the closest possible worlds in which I raise my hand (for example) are ones in which the universe evolves exactly as it actually did until the instant before I raised my hand. In other words, the Strong Thesis under (4) amounts to a denial of the need for an "orderly transition from actual past to counterfactual present and future" (Lewis 1979: 463) in the closest possible worlds. But these 'transition periods' are quirks of Lewis's theory of the closeness ordering of possible worlds, which (as we shall have occasion to discover in Chapter Four) not everyone accepts. In any case, the truth of the Strong Thesis under (4) would *not* confer on us any marvellous abilities like the ability to travel faster than light.

Lewis distinguishes between two interpretations of 'render false', the weak sense and the strong sense. He then argues that we can render the laws false in the weak but not the strong sense. But in doing so, he himself equivocates between two different interpretations of 'miracle'. So Lewis's response to the consequence argument fails. Here, to repeat, is the correct response. An ability to act otherwise amounts to an ability to falsify L&P, if determinism is true. Hence, if I can act otherwise, I can perform an action such that, were I to perform it, either L or P would (have to) be *false*; but it does not follow from this that I can perform an action such that, were I to perform it, either L or P would (have to) be *falsified, whether by my action or by any other event.*

Let me finish this section with a speculative suggestion as to why so many people have gotten into such a muddle in discussing the consequence argument. 'Render false' was introduced as a term of art – van Inwagen was perfectly within his rights to give it whatever meaning he liked. But I suspect he used this particular locution precisely because it already has intuitive natural-language meaning, *and this meaning is context-*

sensitive. I strike a match; the match burns. Did my striking the match 'render false' the proposition that the match is unburnt?[16] In one sense, no – there are possible worlds in which I strike the match and the proposition that the match is unburnt is true, ones in which there is no oxygen in the atmosphere, for example – but in another sense, yes – *holding fixed* various facts, such as the fact that there is oxygen in the atmosphere, the proposition that the match is unburnt is false in every possible world in which my striking of the match occurs. Both the consequence argument and Lewis's response to it seem to work only by exploiting this context-sensitivity, by covertly holding different facts fixed at different points in the argument.

4. The Conditional Analysis

In this section, I will argue that a contextualist treatment of (AP) can be used to make sense of a curious "strategem" (Chisholm: 1964: 28), endorsed by Edwards (1754), Moore (1912), Hobart (1934), Ayer (1954), and according to van Inwagen (1983: 114), "the great majority of the present-day defenders of compatibilism". These philosophers think that the right solution to the freedom and determinism problem is to reject (AP). But they also accept that (AP) seems very plausible.[17] To explain this, these compatibilists argue that, although (AP) is false, a closely related principle is true:

> CONDITIONAL ANALYSIS (CA): An agent's action was free only if she *would* have acted otherwise *had she chosen to act otherwise.*

[16] C.f. Chapter One, Section 1, above.

[17] To paraphrase Lewis (1996: 550) once more: If you are contented in your rejection of (AP), be honest, be naive, hear it afresh. 'I acted freely, yet I couldn't have acted otherwise'. Even if you've numbed your ears, doesn't this *still* sound wrong?

The reason why (AP) seemed so plausible, despite its bizarre implication that freedom is incompatible with determinism, is that we were confusing it with (CA), which does not have this consequence – or so the argument goes.[18]

There is something puzzling about this proposal, on the face of it. 'S could have acted otherwise' is a bare possibility claim of the form ◇P, where P is the proposition that S acted otherwise. 'S would have acted otherwise had she chosen to act otherwise' is a counterfactual of the form (Q □→ P), where Q is the proposition that S chose to act otherwise. On the orthodox semantics of these expressions, ◇P is true if and only if there is an accessible world in which P is true, and (Q □→ P) is true if and only if P is true in every closest accessible world in which Q is true. Now consider a particular modal base **B** and ordering source **O**. Suppose ◇Q is false relative to **B** and **O** – I couldn't have chosen to act otherwise in this context. Then (Q □→ P) is *vacuously* true relative to **B** and **O**, since if there is no accessible world in which Q is true, trivially everything is true in every *closest* such world, including the proposition that S acts otherwise (as well as, for that matter, the proposition that S *doesn't* act otherwise). Suppose then that ◇Q is true relative to **B** and **O** – I could have chosen to act otherwise in this context. Then (CA) actually *implies* (AP) relative to **B** and **O**. If (Q □→ P) is non-vacuously true, then P is true in every closest accessible world in which Q is true; and so there is an accessible world in which P is true; and so ◇P is true. In other words, if S *would* have acted otherwise had she chosen to do so, then she *could have acted otherwise* – period.

The converse doesn't hold, however – from the fact that I could have acted otherwise, it doesn't follow that I *would* have acted otherwise had I so chosen, because P may be true in some accessible world without being true in every closest accessible world in which Q is true. To illustrate this, suppose I made coffee this morning. Now consider a context

[18] Many of these authors go further and claim that all utterances of the form 'S could have acted otherwise' are actually *elliptical* ways of expressing the proposition that S could have acted otherwise had she chosen to act otherwise, a suggestion that Austin (1961) rightly criticises.

relative to which I could have done otherwise – I could have made tea, or indeed nothing at all, instead – but I couldn't have made hot chocolate, because I don't have any in the house. My housemate, however, has saved an old tin of hot chocolate and used it to store her instant coffee. Suppose also that if I hadn't chosen to make coffee, I would have chosen to make hot chocolate (it's my second-favourite breakfast drink). Finally, suppose that I am pretty bad at telling hot chocolate granules and coffee granules apart, especially early in the morning. Then it follows that, if I had chosen *not* to make coffee, I would still have made coffee, since I would have chosen to make hot chocolate but mistaken the coffee granules in the tin for hot chocolate granules. So this is a case in which, relative to a suitable context, I *could* have acted otherwise, even though I *wouldn't* have acted otherwise had I chosen to act otherwise. So there are actions that, relative to a suitable context, are free according to (AP) but not according to (CA).

It follows that (CA) is either vacuously true or *strictly stronger* than (AP), depending on the modal base and ordering source. So long as (CA) has any substantive content at all, any consequence of (AP) is *also* a consequence of (CA). Hence if (AP) implies that freedom and determinism are incompatible, as Edwards, Ayer, and Moore would have us believe, then (CA) *also* has this consequence, or else it's simply vacuous. It follows that replacing (AP) with (CA) is not a satisfactory solution to the freedom and determinism problem.

If this strategy is so obviously fallacious, why was it proposed in the first place? The freedom contextualist can answer this question too. Suppose we're in a demanding context, one relative to which it's incompatible with determinism that I could have acted or chosen to act otherwise. If determinism is true, it's vacuously true relative to this context that I would have acted otherwise had I chosen to act otherwise. But everyday speech is governed by what Lewis (1986) calls the 'rule of accommodation': What one says makes itself non-vacuous, if at all possible, by creating a context relative to which it is non-vacuous. The effect of uttering the sentence 'If I had chosen to act otherwise, I

would have acted otherwise' is often to change the context to one relative to which its content is non-trivial, by expanding the sphere of accessible worlds (equivalently: by weakening the modal base and/or ordering source), to include worlds in which I chose to act otherwise. If it's non-vacuously true relative to this *new* context that I would have acted otherwise had I chosen to act otherwise, then it's *also* true relative to this context that I could have acted otherwise. (CA) seems better suited to compatibilism than (AP), not because it is weaker than (AP), but because charitably interpreting it forces us into a context relative to which the possibility of acting otherwise is compatible with determinism.

Interestingly, the standard response to (CA) from incompatibilists has been to insist – not, as I have done, that (CA) is either vacuous or strictly stronger than (AP), depending on the context – but rather that (CA) is too *weak* (see Chisholm 1964: 26-7; van Inwagen 1983: 114-9). Here's an example due to McKenna and Coats (2015). Suppose Danielle is pathologically afraid of blond dogs. On her sixteenth birthday, her father asks her to choose between two puppies, one a blond-haired Labrador and the other a black-haired Labrador. Danielle picks up the latter. Could she have acted otherwise? "It seems not," McKenna and Coats reply:

> Picking up the blond Lab was an alternative that was not available to her. In this respect, *she could not have done otherwise*...But notice that, *if* she [had] wanted to pick up the blond Lab, *then she would have done so.* (McKenna and Coats 2015; emphasis in original).

McKenna and Coats seem to be suggesting that Danielle's action wasn't free according to (AP), but was free according to (CA). But I have argued that (AP) is strictly weaker than (CA), at least relative to those contexts in which (CA) is non-vacuous. So what's going on here?

Again, the problem here is that different claims are being evaluated relative to different contextual parameters. Here are the two sentences we're interested in:

(5) Danielle could have picked up the blond dog.

(6) Danielle would have picked up the blond dog if she had chosen to do so.

In interpreting (5), the natural modal base to use is one in which Danielle's psychological disorder (together with other salient background conditions) is held fixed, so that there is no accessible world in which Danielle chooses to pick up the blond dog (given determinism), and hence no accessible world in which Danielle picks up the blond dog. Relative to this modal base, then, (5) is false and (6) is vacuously true. The argument seems to work only because principles of charity require us to evaluate (6) relative to a *different* modal base, one which doesn't include facts about Danielle's psychological condition. Relative to *this* modal base, (6) is non-vacuously true, but then so is (5), since there is an accessible world in which Danielle chooses to pick up the blond dog and does just that. There is no modal base relative to which (6) is non-vacuously true and (5) is false;[19] the problem with (CA) is *not* that it is too weak.

Some philosophers (Fara 2008; Vihvelin 2004) have advocated replacing (AP) with a slightly different principle:

DISPOSITIONAL ANALYSIS (DA): An agent acts freely only if she was *able* (had the ability) to act otherwise.

Until recently, (DA) was thought to be equivalent to (CA), since it was assumed that abilities (and dispositions more generally) could be analysed in terms of what agents or objects *would* do in certain circumstances.[20] We've learned the hard way that this is not the case.[21] This has led to renewed attempts to defend (DA) as a solution to the freedom

[19] It doesn't necessarily follow from this that there is no *context* relative to which (5) is false and (6) is non-vacuously true, however, since a single context may supply more than one modal base and ordering source; see also Chapter One, note 16.

[20] This 'simple conditional analysis' of disposition ascriptions is explicitly endorsed by Ryle (1949), Goodman (1954) and Quine (1960), among others.

[21] The first official refutation of the simple conditional analysis is due to Martin (1994), although it had been "a matter of folklore" long before that, according to Lewis (1997: 143).

and determinism problem. But although these authors reject the 'simple conditional analysis' of dispositional claims, they tend to hold on to some more complicated counterfactual account, such as Lewis's 'reformed conditional analysis'.[22] I think this is a mistake. Vetter (2014) has convincingly argued that no counterfactual analysis of dispositional claims can succeed. Instead, Vetter argues, sentences of the form 'S is able to φ' are best interpreted as *bare possibility claims*, analysed in the usual Kratzerian way – S is able to φ relative to **B** and **O** if and only if S φ-s in some element of max(**B**, **O**).[23]

On this view, ascriptions of abilities and dispositions are context-sensitive, in the same way that modal claims are. Suppose Ann, a concert pianist, is currently on a plane.[24] There's one sense in which she is able to play the piano – she has the requisite skills and capacities – and a different sense in which she isn't able to play the piano – there is no piano on the plane. These different interpretations correspond to whether we're 'holding fixed' – i.e. including in the modal base – the fact that Ann is currently nowhere near a piano: If we do hold this fact fixed, there is no accessible world in which Ann is playing the piano, whereas if we don't hold this fact fixed, such a world may well be accessible. Proponents of (DA) do acknowledge the semantic variability of dispositional language, up to a point: Vihvelin (2013), for example, distinguishes between 'wide abilities' and 'narrow abilities', such that Ann has the wide but not the narrow ability to play the piano. But this way of talking is misleading, since it suggests that 'is able to' is *ambiguous* between two different meanings.[25] On the present proposal, 'is able to', just like 'can',

[22] "Something *x* is disposed at time *t* to give response *r* to stimulus *s* iff, for some intrinsic property *B* that *x* has at *t*, for some time *t`* after *t*, if *x* were to undergo stimulus *s* at time *t* and retain property *B* until *t`*, *s* and *x*'s having of *B* would jointly be an *x*-complete cause of *x*'s giving response *r*" (Lewis 1997: 157).

[23] In support of this proposal, note that dispositional adjectives are often most naturally paraphrased using the word 'can': 'elastic', for example, is defined by the *Oxford English Dictionary* as "can be stretched without permanent alteration of size or shape".

[24] This example is due to Franklin (2015).

[25] Indeed, Vihvelin takes her proposal to be inconsistent with van Inwagen's (2008) insistence that "'has the ability' always means the same thing in ordinary English" (Vihvelin 2013: 8).

expresses the same thing on every occasion of use: A function with two hidden argument places for a modal base and an ordering source provided by the conversational context.

Of course, like most modal locutions, 'is able to' plausibly comes with restrictions on the range of allowable modal bases and ordering sources relative to which it can be evaluated.[26] I've bought a lottery ticket; hence there's a sense in which I *could* win the lottery. But there's no obvious sense in which I am *able* to win the lottery. This is true enough, but it doesn't imply that "possibility is not the same as ability" (Vihvelin 2013: 7). Ability is rather a special *kind* of possibility, one defined by a particular set of pairs of modal bases and ordering sources. Call these the *agentive modalities*.[27] Relative to a particular agential modality, then, (DA) and (AP) are equivalent. Insofar as (DA) seems better suited to compatibilism than (AP), it's only because the modalities relative to which (C) is true (those in which the laws of nature and the state of the universe at a time are held fixed) are *not* agentive modalities, so that the construction 'is able to' forces us into a context relative to which (C) is false. Put simply, replacing (AP) with (DA) is not a satisfactory solution to the freedom and determinism problem. The only reason for replacing (AP) with (DA) would be if ascriptions of freedom could *only* be evaluated relative to the agential modalities; and as I argue in the next section, there is in fact very little evidence for such a restriction.

5. Freedom and Norms

In Phillips and Knobe (2009), experimental subjects were presented with the following vignette (adapted from the well-known case in Aristotle's *Nicomachean Ethics*):

[26] C.f. Chapter One, note 10, above.

[27] Maier (2013: 115) reminds us that "the Kratzer semantics alone do not suffice to settle questions about the agentive modalities" – the point is well-taken, and I won't attempt to settle such questions here.

> While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the captain realized that his small vessel was too heavy and the ship would flood if he didn't make it lighter. The only way that the captain could keep the ship from capsizing was to throw his wife's expensive cargo overboard.
>
> Thinking quickly, the captain took her cargo and tossed it into the sea. While the expensive cargo sank to the bottom of the sea, the captain was able to survive the storm and returned home safely. (Phillips and Knobe 2009: 31).

Subjects were then asked whether they agreed or disagreed with the following statement:

> (7) The captain was forced to throw his cargo overboard.

A different group of subjects was presented with a slightly altered story (changes are marked in italics):

> While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the captain realized that his small vessel was too heavy and the ship would flood if he didn't make it lighter. The only way that the captain could keep the ship from capsizing was to throw *his wife* overboard.
>
> Thinking quickly, the captain took *his wife* and tossed *her* into the sea. While the *captain's wife* sank to the bottom of the sea, the captain was able to survive the storm and returned home safely. (Phillips and Knobe 2009: 32).

These subjects were asked whether they agreed or disagreed with the following statement:

> (8) The captain was forced to throw his wife overboard.

As expected, subjects tended to agree with (7) and disagree with (8).[28] This is despite the fact that the only relevant difference between the two cases seems to be a *normative* one

[28] On a scale of 1 to 7, 1 being 'strongly disagree' and 7 being 'strongly agree', (7) received an average score of 4.6 and (8) an average score of 1.9. 4.6 is in the agreement range, but only just, which suggests that some subjects were choosing to interpret the statement relative to an empty

– the captain's actions were morally wrong in the second case but not the first. There is thus "a surprising connection between people's moral judgments and their intuitions about freedom and constraint" (Phillips and Knobe 2009: 32).

Although Phillips and Knobe phrase their target statements in terms of what the captain was *forced* to do, it's plausible to suppose that a similar phenomenon would be observed for subjects' judgements about whether the captain's actions were *free*. (AP) can explain this result (see also Szabó and Knobe 2013). The difference in the judgements elicited in the two cases is due to a difference in the ordering sources made salient by them. In the first case, the context naturally suggests an ordering source relative to which the top-ranked worlds are ones in which the ship *and* its cargo are saved, followed by worlds in which just the ship is saved, worlds in which just the cargo is saved, and finally worlds in which neither is saved. Once we hold fixed the facts about the storm, however, all the worlds in which both the ship and the cargo are saved are eliminated, leaving the worlds in which just the ship is saved tied for first place. So *given* that the captain was in stormy waters, he *had* to throw the cargo overboard, and hence couldn't have acted otherwise than he actually did, relative to this ordering source. So (7), naturally interpreted, is true. In the second case, however, the salient ordering source ranks worlds in which the captain sacrifices himself *at least* as highly as those in which he throws his wife overboard. So even holding fixed facts about the storm, there is an accessible world in which the captain doesn't throw his wife overboard, and so the captain *could* have acted otherwise, relative to this ordering source. So (8), naturally interpreted, is false. That's why subjects agree with (7) and disagree with (8).

Of course, one can interpret these same statements relative to different sets of accessible worlds. Suppose Julie, Aziz, and Desai are discussing the first case:

---

ordering source, as a claim about whether the captain had the *ability* to refrain from throwing the cargo overboard, for example – see below.

Julie: The captain was forced to throw the cargo overboard, so his action wasn't free.

Aziz: Well, the captain wasn't *really* forced to do anything – he *could* simply have gone down with the ship instead – so his action, strictly speaking, *was* free.

Desai: Well, if we're really being picky, no-one can do otherwise than what the laws of nature and the initial state of the universe determine they will do, so actually Julie is right – the captain's action wasn't free.

On the contextualist view defended here, Julie, Aziz and Desai don't really disagree (or at best their disagreement is merely *metalinguistic*, a disagreement about the right way of interpreting (7) in the context). Aziz's utterance changes the context to one with an empty ordering source, and Desai's utterance changes the context again to one with a modal base that includes the initial state of the universe and the laws of nature. Each utterance is true in the context it creates.

Now compare the following two sentences:

(9) The captain could have left the cargo on board.

(10) The captain had the ability to leave the cargo on board.

I've argued that there is a natural context relative to which (9) is false in the first case. The captain *couldn't* have left the cargo on board, because, given that a storm was raging, he *had* to – had no choice *but* to – throw it overboard to save the ship. But I find it hard to hear a similar false reading of (10). The captain was clearly *able* to do nothing and leave the cargo on board, even if exercising this ability wasn't a reasonable option in the circumstances. This suggests that ascriptions of freedom can be evaluated relative to non-agential modalities. If that's right, van Inwagen is wrong when he claims that "[a]ll the phrases that have been used in definitions of 'free will'…can be defined in terms of, or dispensed with in favor of, 'able'" (van Inwagen 2008: 333). The range of allowable interpretations of 'free' outstrip the range of allowable interpretations of 'is able to'.

6. Freedom and Responsibility

Lurking in the background of the freedom and determinism debate, of course, is the following principle:

> MORAL RESPONSIBILITY (MR): An agent is morally responsible for performing an action only if that action was free.

If it turned out that no-one ever acts freely, (MR) would imply that no-one is morally responsible for anything they do. Our whole practice of praise and blame, of reward and punishment, of excuse and justification, would be based on a falsehood (to put it somewhat melodramatically). This is part of the reason why so many philosophers have devoted so much time and intellectual resources to solving the freedom and determinism problem.

On the view defended here, the right-hand-side of (MR) is context-sensitive. Although there are some contexts relative to which an action is 'free' only if determinism is false, there are others relative to which acting 'freely' is perfectly compatible with determinism. But this raises a different kind of problem; for ascriptions of moral responsibility do *not* seem to be context-sensitive.[29] If I utter the sentence 'Kim is tall' and you, in a different context, utter the sentence 'Kim is not tall', it's not hard to imagine a situation in which we don't really disagree – perhaps my utterance expressed the proposition that Kim is tall *for a person*, for example, whereas yours expressed the proposition that Kim is not tall *for a basketball player*. But if I insist that Zara is *morally responsible* for robbing

[29] Some authors have flirted with a contextualist semantics for ascriptions of moral responsibility. Hawthorne (2001: 70) asks whether "contextualism about moral accountability [is] unthinkable" and gives an ambivalent answer. Rieber (2006: 241) considers combining "contextualism about free will with contextualism about moral responsibility", suggesting that "[t]he two theories are attractive in some of the same ways". More explicitly, Baumann (2011: 219) argues, albeit for somewhat different reasons, that "[t]he truth conditions of judgments of the form 'S is morally responsible for x' depend on and vary with the context of the attributor". And Slater (2005: 130) responds to a version of the consequence argument with the suggestion that "[c]laims of responsibility, like claims to knowledge, are context-dependent". For the reasons given below, however, I don't think that embracing a contextualist semantics for ascriptions of moral responsibility is the right solution.

the bank (and hence that she did something *wrong*, for which she *ought to be punished*, that this would be the *just* reaction, and so on), it's absurd to think that there could be a context in which you utter the sentence 'Zara is not morally responsible for robbing the bank' without thereby disagreeing with me. As a number of authors have noted, it seems to be a presupposition of our moral practice that moral terms mean the same thing in every context.[30]

Given the strong conceptual connections between freedom and moral responsibility, how can we stop the context-sensitivity of freedom-talk from infecting the semantics of moral language? I think that this problem is not as serious as it first appears. To see this, note first that many forms of contextualism will face a version of the same challenge. Consider the following (slightly morbid) example due to Slater (2005), for example:

> I am to a great extent off the hook (considerations of moral luck or negligence aside) if I *unknowingly* back over my neighbor's kitten in my car. I am blameworthy, on the other hand, if I did know that it was sleeping under my tire and backed up anyway. (Slater 2005: 129).

Suppose I had lots of evidence that my neighbour's kitten was under my tyre – I saw a cat-like figure as I approached my car, heard a cat-like mewl as I turned on the engine, I ignored several shouted warnings from onlookers as I started reversing, and so on. According to the sceptic, I nevertheless failed to *know* that the kitten was under my tyre, since there are possibilities consistent with my evidence – think "CIA plots, hallucinogens in the tap water, conspiracies to deceive, old Nick himself" (Lewis 1996: 549), and so on – in which the kitten was not under my tyre. Thus the sceptic seems

[30] See Horgan and Timmons (1991), for example. Smith (1994: 35) argues any plausible semantics of moral terms like 'right' and 'wrong' "must make sure that moral claims do not turn out to have different contents in different contexts", since this would be "contrary to the platitude that if A says 'x is right' and B says 'x is not right' then A and B disagree". This form of argument has come under criticism in recent years, however (see Plunkett and Sundel 2013; Björnsson and Finlay 2010; Khoo and Knobe forthcoming). In any case, I don't want to argue that no version of moral contextualism is tenable, but only that my version of freedom contextualism is not committed to it.

forced to conclude that I am not morally responsible for running over the kitten. But the epistemic contextualist admits that there are *some* contexts relative to which I didn't 'know' that the kitten was under my tyre and some contexts relative to which I *did* 'know' this. Is the epistemic contextualist therefore committed to saying that ascriptions of moral responsibility are also context-sensitive?

I think not. The important point is that, in order to be morally responsible for the kitten's unfortunate demise, there is a certain epistemic standard I must have met with respect to the proposition that the kitten was under my tyre. According to the epistemic contextualist, the word 'know', in *some* contexts, expresses the relation which is instantiated if and only if I meet this standard; whereas in other contexts, 'know' expresses a different relation, one governed by different epistemic standards. Intuitively, my epistemic position with respect to the proposition that the kitten was under my tyre in the case above was clearly sufficient to hold me morally responsible for running over the kitten, *regardless* of whether it is correct in the present context to describe this as an instance of 'knowledge'. To put it crudely, there are many different 'kinds' of knowledge, according to the contextualist. But there is only one kind of moral responsibility. What Slater's case illustrates is just that "the "kind" of knowledge required for moral responsibility falls short of the requirements assumed by the skeptic" (Slater 2005: 129).

By analogy with a criticism of epistemic contextualism due to Hawthorne (2003), one might argue that this approach implies the truth, in some contexts, of conjunctions which "seem strange" and "sound odd" (Hawthorne 2003: 87). For example, the contextualist seems forced to accept that there are contexts relative to which the following sentence is true:

> (11) I am morally responsible for the kitten's death, even though I didn't know that the kitten was under my tyre.

(11) does indeed sound odd. But the contextualist can explain its oddness (c.f. DeRose 2009: 246). The idea is that (11) sounds bad, despite the fact that there are contexts relative to which it expresses something true, because it is difficult, if not impossible, to utter (11) without thereby *changing* the context to one relative to which it expresses something false. In other words, the very act of evaluating a subject's moral responsibility for his actions forces us into that very context relative to which 'know' expresses the relation of relevance to moral responsibility, that relation which is required to hold between me and the proposition that the kitten was under my tyre in order for me to be morally responsible for the kitten's death. Although there are epistemic standards relative to which (11) is true, it cannot be truly uttered – and *that's* why it sounds odd.

I think something similar is true about ascriptions of freedom. On my view, an ascription of freedom expresses different propositions in different contexts. In some contexts, it predicates that relation which is required to hold between an agent and an action for the one to be morally responsible for the other; in other contexts, it predicates a different relation. To put it crudely, there are many 'kinds' of freedom, according to the contextualist. But there is only one kind of moral responsibility. The challenge is therefore to determine *which* 'kind' of freedom is the one that is required for moral responsibility.[31]

Framing the issue in this way represents an important shift in the traditional dialectic. Although many philosophers working on the freedom and determinism problem acknowledge that ascriptions of freedom express different things in different contexts, the usual reaction to this is to simply stipulate that one is talking about "the kind of freedom relevant to moral responsibility" (Sartorio 2015: 94), and to then go on to ask whether this kind of freedom requires the existence of 'alternative possibilities' or the falsity of determinism. But these questions ought really to be secondary to the question

[31] C.f. Rieber (2006: 242), who notes that his "contextualist account of free will is consistent with, but does not entail, contextualism about moral responsibility".

of which kind of freedom *is* the one relevant to moral responsibility; and the right way to answer *this* question is by doing ethics, not metaphysics, by examining our concept of moral responsibility, our blaming and praising practices, the norms governing excusing and excepting conditions, and so on.

Consider, for example, the literature on 'Frankfurt-style cases'.[32] Here's an example of such a case due to Fischer (2008). Black, a neurosurgeon and committed Democrat, has secretly inserted an electronic device in Jones's brain. The device continuously monitors Jones's neural activity; should it detect neural patterns indicative of a desire to vote Republican, the device will stimulate Jones's neurons in such a way as to ensure that he chooses instead to vote Democrat. In fact, Jones shows no inclination towards voting Republican; he decides to vote Democrat, and does so, and all the while the device lies dormant. Intuitively, Jones is morally responsible for voting Democrat. Yet there's also an obvious sense in which Jones couldn't have acted otherwise – the device would have ensured that Jones had voted Democrat, whatever his political inclinations. Thus, Frankfurt concludes, the conjunction of (AP) and (MR) is false: "A person may well be morally responsible for what he has done even though he could not have done otherwise" (Frankfurt 1969: 1).

It's easy to see the appeal of Frankfurt-style cases for compatibilists. Rather than "getting entangled in dialectical stalemates" (Fischer 2008: 317) over the right way to solve the freedom and determinism problem, Frankfurt-style cases seem to offer an *independent* route to denying the conjunction of (AP) and (MR). "[T]he moral of Frankfurt stories", according to Fischer (2006: 198), is that they "suggest that alternative possibilities are irrelevant to the grounding of moral responsibility". "[I]f causal determinism rules out

[32] These cases are named in honour of Frankfurt (1969), but were arguably first discussed by John Locke. "[S]uppose a man be carried, whilst fast asleep, into a room where is a person he longs to see and speak with; and be there locked fast in, beyond his power to get out: he awakes, and is glad to find himself in so desirable company, which he stays willingly in, i.e. prefers his stay to going away. I ask, is not this stay voluntary? I think nobody will doubt it: and yet, being locked fast in, it is evident he is not at liberty not to stay, he has not freedom to be gone." (Locke 1689: 153).

moral responsibility", therefore, "it is not in virtue of eliminating alternative possibilities" (Fischer 2008: 318).

The contextualist about ascriptions of freedom offers a very different assessment of the significance of Frankfurt-style cases. What they certainly *don't* show is that (AP) is false. There's a sense in which Jones couldn't have acted otherwise in the case above, and hence didn't act freely, one where we hold fixed the fact that Jones had a device maliciously implanted in his brain; but there is also a sense in which he *could* have acted otherwise, and hence *did* act freely, since there is obviously *some* possible world in which Jones doesn't have a device in his brain and votes Republican. Put this way, it's hard to see how anyone could sensibly deny that there is *some* interpretation of 'could have acted otherwise' according to which an agent is morally responsible for an action only if she could have acted otherwise. Indeed, as Franklin (2015) points out, Fischer's own analysis of moral responsibility implies as much, since according to Fischer, an agent is morally responsible for an action only if the mechanism that produced the action is "appropriately reasons-responsive" (Fischer and Ravizza 1998: 62), where a mechanism is appropriately reasons-responsive if and only if there is "*some* possible world in which there is a sufficient reason to do otherwise, the agent's actual mechanism operates, *and the agent does otherwise*" (Fischer 2006: 68; second emphasis is mine).

The issue at the heart of Frankfurt-style cases, therefore, is not whether moral responsibility requires "the genuine metaphysical availability of alternative possibilities" (Fischer 2012: 186), whatever exactly that means. Modal talk and ascriptions of freedom quantify over different sets of possible worlds, depending on the conversational context. The real issue is what is the *right* context to use for the purposes of evaluating an agent's moral responsibility for her actions. In the case above, it's fairly obvious that Jones should be held morally responsible for voting Democrat. We learn, therefore, that the context relative to which 'free' expresses the relation of relevance to moral responsibility is one in which certain kinds of facts (facts about electronic devices in brains, for

example) are not included the modal base. By considering other similar cases, and examining our intuitions about them, we might learn more facts of this kind. But we will *not* succeed in establishing that (AP) is false.

Of course, this whole project would be a phenomenal waste of time if it turned out that agents must act freely *given the initial state of the universe and the laws of nature* in order to be held morally responsible for their actions. *Then* it would follow, if determinism is true, that no-one is morally responsible for anything they do, regardless of whether they have electronic devices in their brains.[33] But this would be some error theory indeed. It goes against even our most basic beliefs about praise and blame. The burden of proof is therefore on the sceptic to show why we should take such a view seriously. And it's *not* enough, in discharging this burden, to point out that there are contexts in which it seems true to say that an agent could have acted otherwise only if determinism is false. There *are* such contexts, but there are many more permissive contexts too, relative to which ascriptions of freedom express different propositions, ones which are perfectly compatible with the falsity of determinism. The sceptic must show that it is the *demanding* contexts that are relevant to moral responsibility; and as far as I can see, there is no reason to think that this is the case, and every reason to think that it isn't.

Our intuitions in Frankfurt-style cases are fairly robust. But other cases are not so clear-cut. In the next section, I will consider what I think is a more interesting example, one

[33] C.f. Goetz (2005: 320-1), who argues that Frankfurt-style cases simply beg the question against the incompatibilist, since "without the obtaining of causal determinism in the actual sequence of events, the device cannot prevent Jones from making an alternative choice, and with causal determinism in the actual sequence of events it is not the device that prevents Jones from making an alternative choice". In other words, if determinism is false, the device in Jones's brain is impotent; and if determinism is true, Jones couldn't have acted otherwise, device or no device. Either way, the device is irrelevant. See Fischer (2010) for discussion and a reply.

which has received comparatively little attention from philosophers[34] – the debate over the common law defence of *duress*.

7. Duress: A Justification or an Excuse?

There are two elements to every crime – the *mens rea* ('guilty mind') and the *actus reus* ('guilty act'). Murder, for example, consists of a *mens rea* of intent to kill or inflict grievous bodily harm and an *actus reus* of unlawful killing. But even if it can be proved beyond reasonable doubt that a defendant satisfies both these elements, she may still escape conviction by appealing to one of a number of legal defences. Some defences, like the statute of limitations, entrapment or infancy, are explicitly motivated by considerations of public policy. But the others can be divided into two categories – the *justifications* and the *excuses*.

According to the US Model Penal Code, "[t]o say that someone's conduct is 'justified' ordinarily connotes that the conduct is thought to be right, or at least not undesirable".[35] Suppose D drives his son to work, despite being disqualified from driving. D satisfies the elements of a road traffic offence. But suppose also that D's wife had threatened to commit suicide if D didn't drive their son to work.[36] Then D may appeal to the legal defence of *necessity* to escape conviction. In this case, D arguably had a moral *duty* to break the law. But justifications may also apply when the defendant merely had a moral *right* to break the law. Suppose, for example, that D intentionally hits V. Ordinarily, D would be guilty of an offence against the person. But if D had agreed with V to enter a boxing ring for a few rounds, D would be able to appeal to the legal defence of *consent* to

[34] Hyman (2015) is an excellent exception.

[35] *Model Penal Code Commentaries*, art. 3, introduction, 2-3.

[36] *R v. Martin* [1989] 88 Cr App R 343.

escape conviction. V's agreement afforded D a moral right, if not a moral duty, to hit V, so D's conduct was justified.[37]

"[T]o say that someone's conduct is 'excused'", by contrast, "ordinarily connotes that the conduct is thought to be undesirable but that for some reason the actor is not to be blamed for it".[38] Suppose D intentionally kills V, thereby satisfying the elements of murder. But suppose also that D suffers from a form of psychosis that leads him to believe that God has commanded him to kill V. Then D may appeal to the legal defence of *insanity* to escape conviction.[39] In allowing the defence, we do not suggest that we *approve* of D's actions, or that D, in the circumstances, had a moral *right* to do what he did. Rather, we express the opinion that D cannot be held morally responsible for his actions, wrongful though they were.

Most non-procedural defences can be neatly classified either as justifications or excuses, with one notorious exception – the common law defence of *duress*. Duress applies to those crimes committed under unavoidable threat of immediate death or grievous bodily harm. Suppose D intentionally steals money from a bank. D satisfies the elements of theft. But suppose P had threatened to break D's legs unless he carried out the theft; then

---

[37] While it is the responsibility of the prosecutor to prove beyond reasonable doubt that the defendant satisfies the elements of a crime, it is the responsibility of the defendant to prove beyond reasonable doubt that she has a defence; and typically, the standards of proof for justifications like necessity tend to be fairly high. There are good Rule of Law reasons for this – if a person were able to escape criminal conviction whenever her conduct was morally justified, the law would offer little guidance on how to behave, beyond 'Don't do what you don't have a moral right to do'. As a result, necessity tends to only be allowed in exceptional cases where failing to commit the crime would have an egregious moral wrong. As Lord Justice Davies puts it:

> The law regards with the deepest suspicion any remedies of self-help, and permits these remedies to be resorted to only in very special circumstances. The reason for such circumspection is clear – necessity can very easily become simply a mask for anarchy. (*Southwark London Borough Council v. Williams* [1971] 1Ch 734, 846).

[38] *Model Penal Code Commentaries*, art. 3, introduction, 2-3.

[39] According to the M'Naughten rules (*Queen v. M'Naghten*, 8 Eng. Rep. 718 [1843]), insanity should be allowed as a defence in cases where, "because of mental disease, [the defendant] did not know that what he was doing was wrong, [but] believed, for example, that he was carrying out a command from God" (*Price v. Commonwealth,* 228 Va. 452, 460).

D may appeal to duress to escape conviction. The dilemma about duress is this: Is duress a justification or is it an excuse?

The orthodox view, in most jurisdictions at least, is that duress is properly conceptualised as an excuse; indeed "the quintessential excuse" (Simester *et al.* 2013: 739). The reasoning here is something like this. If D robs a bank under threat of grievous bodily harm, it doesn't seem right to say that the threat afforded D a moral right (never mind placed him under a moral duty) to rob the bank. Indeed, had D refused to obey P and instead submitted to the threatened harm, we would normally make him the subject of *praise* for his heroic act of self-sacrifice. Duress, therefore, is not a justification but an excuse, "more the product of a concession to human frailty than of a societal determination that the actor's conduct was, on balance, not wrongful" (Chiesa 2008: 752).

The orthodox analysis has its detractors, however. Some have defended the opposite view, that duress is a justification, not an excuse (see Westen and Mangiafico 2003). When D robs a bank under threat of serious harm, he is fully responsible for his actions, on this view. But what he did was, morally speaking, the right thing to do in the circumstances. After all, the physical harm D would have experienced had he submitted to the threat is greater than the merely financial loss the bank in fact experienced. Hence the action D chose to perform was the one which maximized utility, and therefore the morally correct one, regardless of the fact that D's actions were designed to avoid harm to himself.[40] Had D chosen instead to submit to the threatened harm out of some misplaced sense of heroism, we might excuse his conduct, even admire his character; but we would nevertheless have to concede that what he did was wrong.

[40] "The rationale for the defence of duress is that, for reasons of social policy, it is better that the defendant, faced with a choice of evils, choose to do the lesser evil (violate the criminal law) in order to avoid the greater evil threatened by the other person." (LaFave and Scott 1986: 433).

There are good arguments on both sides of this debate. Supporters of the duress-as-justification view argue that their opponents cannot explain the fact that appeals to duress tend not be allowed unless the severity of the harm threatened exceeds that of the harm caused (and *a fortiori*, not allowed as a defence to murder or manslaughter).[41] After all, if a person threatened with death cannot be held responsible for his actions, then he ought to be excused *regardless* of the harm caused by those actions. In response, supporters of the duress-as-excuse view simply reject the legitimacy of this restriction – "If duress is an excuse, the common law blanket rule excluding it as a defence to murder should be rejected" (Chiesa 2008: 753). On the other hand, supporters of the duress-as-excuse view argue that their opponents cannot explain the fact that appeals to duress tend not to be allowed unless the harm threatened is death or serious bodily injury. After all, the difference in utility between a £2.40 loss in revenue and a broken finger is arguably greater than the difference in utility between a broken arm and a broken leg, but whereas D can appeal to duress to defeat a charge of grievous bodily harm if he broke V's arm under threat of a broken leg, he cannot appeal to duress to defeat a penalty fare if he boarded a London Underground train without a ticket under threat of a broken finger. In response, supporters of the duress-as-justification view simply reject the legitimacy of this restriction – "It is not proper, on principle, to limit the defense of duress to situations where the instrument of coercion is a threat of death or serious bodily injury" (LaFave and Scott 1986: 438).

My aim here is not to come down on one side of the debate or the other. But there is one important objection that is often levelled at the duress-as-excuse view, which I think is importantly mistaken in light of the contextualist treatment of (AP) defended here. The standard explanation of the excuses is that they reflect "a fundamental principle of

[41] *Abbott v. The Queen* 1977 AC 755.

morality that a person is not to be blamed for what he has done if he could not help doing it" (Hart 1968: 152):

> What is crucial is that those whom we punish should have had, when they acted, the normal capacities, physical and mental, for abstaining from what it [the law] forbids…Where these capacities…are absent…the moral protest is that it is morally wrong to punish because 'he could not have helped it' or 'he could not have done otherwise' or 'he had no real choice' (Hart 1968: 152).

The excuses, in other words, are usually explained by appeal to (MR) and (AP), which together imply that an agent is morally responsible for an action only if she could have acted otherwise. Someone with psychosis, for example, is excused because her actions weren't free, and they weren't free because she couldn't have acted otherwise. But it's widely believed that this explanation is *not* available in the case of duress. Unlike the criminally insane or the psychologically incapacitated, those who act under duress often do so calmly and deliberately, after a thorough assessment of the costs and benefits of compliance. These people clearly *could* have acted otherwise, the argument goes; it's just that acting otherwise wasn't an attractive option in the circumstances. "[H]owever desirable it may be to avert a threat, it is not obvious that this is something a person subjected to duress *must* do", according to Hyman (2015: 82). As Hart and Honoré (1985: 157) put it, "[t]he person threatened is literally able to choose whether to act as instructed or suffer the threatened harm, so that a choice exists". Lord Simon summarizes the position as follows:

> [In cases of duress,] there is power of choice between two alternatives, but one of the alternatives is so disagreeable that even serious infraction of the criminal law seems preferable…duress is not inconsistent with act and will, the will being deflected, not destroyed.[42]

[42] *DPP for N. Ireland v. Lynch* [1975] AC 653, 692ff.

Notwithstanding this consensus, however, I think there *is* a salient sense in which an agent who acts under duress couldn't have acted otherwise, and hence did not act freely. There are two salient interpretations of 'could have done otherwise' in the case of the bankrobber, for example. One corresponds to a modal base that includes the fact that D will suffer serious bodily harm if he doesn't rob the bank (together with other salient background facts) and an *empty* ordering source. Relative to this context, there are accessible worlds in which D robs the bank and accessible worlds in which he doesn't (and therefore suffers the threatened harm). This is the sense in which D had a choice – he didn't *have* to rob the bank; he *could* have simply submitted to the threatened harm. But now consider the same case relative to an ordering source which ranks these worlds in light of D's interests to bodily security. D robs the bank in all of the top-ranked worlds on this ordering. This is the sense in which D *didn't* have a choice – he *had* to rob the bank; he *couldn't* have simply refused. The right question to ask is not, 'But was D's action *really free*?'[43] – one cannot speak of an action's being free independently of a set of accessible worlds. Rather, the right question to ask is this: Is the distinctive *kind* of freedom lacked by those who act under duress – freedom *in light of* interests to bodily security, *given* a threat of serious harm – required for moral responsibility?

The right way to answer this question, to repeat, is by doing ethics, not metaphysics. If we believe that D acts wrongfully, but blamelessly, in robbing the bank under threat of serious bodily harm, we should conclude that the distinctive kind of freedom in question *is* necessary for moral responsibility. Duress would then be explained in exactly the same way as the other excuses, by appeal to (MR) and (AP). If, on the other hand, we believe that D is morally *praiseworthy* for robbing the bank under threat of serious bodily harm, we should conclude that the distinctive kind of freedom in question is *not* necessary for

[43] As Slater points out, adding '*really*' before a context-sensitive term often has the effect of raising the standards for its application – "If 'France is hexagonal' is reckoned true according to some perspective (from very high up) and set of standards, asking (in a certain tone of voice, perhaps), 'but is France *really* hexagonal?' will no doubt get us to revise our earlier reckoning. It does so by changing the context of the evaluation." (Slater 2005: 116).

moral responsibility. Duress would then have to be classified as a justification, rather than as an excuse. It should really go without saying, however, that at no point during the course of this debate would it be necessary (or indeed appropriate) to consider whether the fundamental physical laws are deterministic.

8. Conclusion

This paper has defended a contextualist dissolution of the freedom and determinism problem, one which combines the ancient idea that free action requires the possibility of acting otherwise with Kratzer's semantics of modality. On this view, although there are some contexts relative to which an agent acts freely only if determinism is false, there are many other contexts relative to which this is not the case. We can thus preserve the intuition that many everyday ascriptions of freedom are *true*, without rejecting (AP). I've also described a number of other virtues of the contextualist proposal – it can explain what's wrong with the consequence argument, it can explain what's wrong with the conditional analysis, it can explain why ascriptions of freedom are sensitive to normative facts, and it can clarify the relationship between freedom and moral responsibility. This, I submit, is good evidence that the freedom and determinism problem is in fact no problem at all.

References

Arregui, A. (ms). Layering modalities: The case of backtracking conditionals.

Austin, J. L. (1961). Ifs and cans. In J. L. Austin, *Philosophical papers*, Oxford: Clarendon Press.

Ayer, A. J. (1954). Freedom and necessity. In A. J. Ayer, *Philosophical essays*, London: Macmillan and Company.

Balaguer, M. (2002). A coherent, naturalistic, and plausible formulation of libertarian free will. *Noûs 36*(3), 379-406.

Baumann, P. (2011). A puzzle about responsibility. *Erkenntnis 74*(2), 207-224.

Beebee, H. (2003). Local miracle compatibilism. *Noûs 37*(2), 258-277.

Björnsson, G. and Finlay, S. (2010). Metaethical contextualism defended. *Ethics 121*(1), 7-36.

Chiesa, L. E. (2008). Duress, demanding heroism and proportionality. *Vanderbilt Journal Transnational Law 41*, 741-773.

Chisholm, R. (1964[2003]). Human freedom and the self. In G. Watson (ed.), *Free will*, Oxford: Oxford University Press.

Clarke, R. (2003). *Libertarian accounts of free will*. New York: Oxford University Press.

DeRose, K. (2009). *The case for contextualism: Knowledge, scepticism, and context (vol. 1)*. Oxford: Oxford University Press.

Earman, J. (1986). *A primer on determinism*. Dortrecht: D. Reidel.

Edwards, J. (1754[1957]). *Freedom of will (ed. P. Ramsey)*. New Haven: Yale University Press.

Fara, M. (2008). Masked abilities and compatibilism. *Mind 117*(468), 843-865.

Fischer, J. M. (2006). *My way: Essays on moral responsibility*. New York: Oxford University Press.

Fischer, J. M. (2008). Responsibility and the kinds of freedom. *The Journal of Ethics 12*(3-4), 203-228.

Fischer, J. M. (2010). The Frankfurt cases: The moral of the stories. *Philosophical Review 119*(3), 315-336.

Fischer, J. M. (2012). *Deep control: Essays on free will and value*. New York: Oxford University Press.

Fischer, J. M. and Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.

Frankfurt, H. (1969). Moral responsibility and the principle of alternative possibilities. *The Journal of Philosophy 66*(24), 829-839.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy 68*(1), 5-20.

Franklin, C. E. (2015). Everyone thinks that an ability to do otherwise is necessary for free will and moral responsibility. *Philosophical Studies 172*(8), 2091-2107.

Goetz, S. C. (2005). Frankfurt-style counterexamples and begging the question. *Midwest Studies in Philosophy 29*(1), 83-105.

Goodman, N. (1954). *Fact, fiction and forecast*. Cambridge, MA: Harvard University Press.

Hart, H. L. A. (1968). Punishment and the elimination of responsibility. In H. L. A. Hart, *Punishment and responsibility*, Oxford: Oxford University Press.

Hart, H. L. A. and Honoré, T. (1985). *Causation in the law (2nd ed.)*. Oxford: Clarendon Press.

Hawthorne, J. (2001). Freedom in context. *Philosophical Studies 104*(1), 63-79.

Hawthorne, J. (2003). *Knowledge and lotteries*. Oxford: Oxford University Press.

Hobart, R. E. (1934). Free will as involving determinism and inconceivable without it. *Mind 43*(169), 1-27.

Hobbes, T. (1654[1750]). Of liberty and necessity. In T. Hobbes, *The moral and political works of Thomas Hobbes of Malesbury*, London.

Hoefer, C. (2016). Causal determinism. In E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy (Spring 2016 Edition)*, URL = <http://plato.stanford.edu/archives/spr2016/entries/determinism-causal/>.

d'Holbach, P. T. (1770[1999]). *The system of nature.* Manchester: Cilinamen Press.

Horgan, T. and Timmons, M. (1991). New wave moral realism meets moral twin earth. *Journal of Philosophical Research 16*, 447-465.

Hume, D. (1739[1975]). *A treatise of human nature (ed. P. H. Nidditch).* Oxford: Clarendon Press.

Hume, D. (1748[2006]). *An enquiry concerning human understanding (ed. T. L. Beauchamp*). Oxford: Clarendon Press.

Hyman, J. (2015). *Action, knowledge and will.* Oxford: Oxford University Press.

van Inwagen, P. (1983). *An essay on free will.* Oxford: Clarendon Press.

van Inwagen, P. (2008). How to think about the problem of free will. *The Journal of Ethics 12*(3), 327-341.

Kane, R. (1985). *Free will and values.* Albany: State University of New York Press.

Kane, R. (1996). *The significance of free will*. New York: Oxford University Press.

Khoo, J. and Knobe, J. (forthcoming). Moral disagreement and moral semantics. *Noûs*.

Kratzer, A. (1977). What 'must' and 'can' must and can mean. *Linguistics and Philosophy 1*(3), 337-355.

Kratzer, A. (1981). The notional category of modality. In H. J. Eikmeyer and H. Rieser (eds.), *Words, worlds and contexts*, New York: de Gruyter.

LaFave, W. R. and Scott, A. W. (1986). *Criminal law*. West Publishing Company

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs 13*(4), 455-476.

Lewis, D. (1981). Are we free to break the laws? *Theoria 47*(3), 113-121.

Lewis, D. (1986). Scorekeeping in a language game. *Journal of Philosophical Logic 8*(1), 339-359.

Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy 74*(4), 549-567.

Lewis, D. (1997). Finkish dispositions. *The Philosophical Quarterly 47*(187), 143-158.

List, C. (2014). Free will, determinism, and the possibility of doing otherwise. *Noûs 48*(1), 156-178.

Locke, J. (1689[1836]). *An essay concerning human understanding*. London: T. Tegg and Son.

Maier, J. (2013). The agentive modalities. *Philosophy and Phenomenological Research 87*(3), 113-134.

Martin, C. B. (1994). Dispositions and conditionals. *The Philosophical Quarterly 44*(174), 1-8.

McKay, T. and Johnson, D. (1996). A reconsideration of the argument against compatibilism. *Philosophical Topics 24*(2), 113-122.

McKenna, M. and Coates, D. J. (2015). Compatibilism. In E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy (Summer 2015 Edition)*, URL = <http://plato.stanford.edu/archives/sum2015/entries/compatibilism/>.

Moore, G. E. (1912). *Ethics.* London: Williams and Norgate.

Nowell-Smith, P. (1948). Freewill and moral responsibility. *Mind* 57(225), 45-61.

Pereboom, D. (2001). *Living without free will.* Cambridge: Cambridge University Press.

Phillips, J. and Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry 20*(1), 30-36.

Plunkett, D. and Sundell, T. (2013). Disagreement and the semantics of normative and evaluative terms. *Philosophers' Imprint 13*(23).

Quine, W. V. (1960). *Word and object.* Cambridge, MA: The MIT Press.

Rieber, S. (2006). Free will and contextualism. *Philosophical Studies 129*(2), 223-252.

Ryle, G. (1949). *The concept of mind*. London: Penguin.

Sartorio, C. (2015). The problem of determinism and free will is not the problem of determinism and free will. In A. R. Mele (ed.), *Surrounding free will: Philosophy, psychology and neuroscience*, Oxford: Oxford University Press.

Simester, A. P., Sullivan, G. R., *et al.* (2013). *Simester and Sullivan's criminal law: Theory and doctrine*. Oxford: Hart Publishing.

Slater, M. H. (2005). A contextualist reply to the direct argument. *Philosophical Studies 125*(1), 115-137.

Smith, M. (1994). *The moral problem*. Cambridge: Blackwell.

Szabó, Z. G. and Knobe, J. (2013). Modals with a taste of the deontic. *Semantics and Pragmatics 6*(1), 1-42.

Vetter, B. (2014). Dispositions without conditionals. *Mind 123*(489), 129-156.

Vetter, B. (2015). *Potentiality: From dispositions to modality*. Oxford: Oxford University Press.

Vihvelin, K. (2004). Free will demystified: A dispositional account. *Philosophical Topics 32*(1-2), 427-450.

Vihvelin, K. (2013). *Causes, laws, and free will*. New York: Oxford University Press.

Westen, P. and Mangiafico, J. (2003). The criminal defense of duress: A justification, not an excuse – and why it matters. *Buffalo Criminal Law Review 6*(2), 833-950.

Wiggins, D. (1973). Towards a reasonable libertarianism. In T. Honderich (ed.), *Essays on freedom of action,* London: Routledge and Kegan Paul.

Chapter Three

Causes and Counterparts

Abstract: It follows from David Lewis's counterpart-theoretic approach to *de re* modality and his counterfactual theory of causation that causal claims are relativized to a set of counterpart relations. Call this Shlewis's view. I show how Shlewis's view can provide attractively unified solutions to similar modal and causal puzzles. Although I don't endorse Shlewis's view, I argue that it is better motivated, by his own lights, than the view Lewis actually held, and also better motivated than a similar approach which relativizes causal claims to sets of 'contrast events'.

1. Introduction

In his *Events*, David Lewis (1986a) considers the following case. Suppose John says 'Hello' to Fred. He says it rather too loudly. But it wasn't John's saying 'Hello' *loudly* that caused Fred to greet him in return; rather, it was John's saying 'Hello' that caused Fred to greet him in return. For suppose John's saying 'Hello' loudly hadn't occurred – then John would still have said 'Hello', only not so loudly, and so Fred would still have greeted John in return.

Lewis concludes from this that 'John's saying "Hello"' and 'John's saying "Hello" loudly' denote distinct events. The former event is essentially a saying-'Hello' but only accidentally loud; it would still have occurred if John had spoken softly. The latter is essentially a saying-'Hello'-loudly, and it would not have occurred if John had said 'Hello' but said it softly. In virtue of differing in their modal properties, the two events also differ

in their causal effects – the former caused Fred to greet John in return, whereas the latter did not.[1]

Lewis grants that "[t]here is a persuasive intuition…that it is wrong to count both the first and the second event because if we do, we count something twice over" (Lewis 1986a: 256). But he seems willing to abandon this intuition in order to account for our causal judgements. In this paper, I describe a different possible response to this example. It requires only two ingredients: Lewis's modal counterpart theory and Lewis's counterfactual account of causation. Put these together and it follows that causal claims are relativized to a choice of counterpart relations, determined by the conversational context. Assuming that one of the relevant features of the context is one's choice of event description, we can explain the different semantic roles of 'John's saying "Hello"' and 'John's saying "Hello" loudly' in causal contexts, without giving up the intuitive idea that they denote one and the same event. For reasons that will become clearer below, I call this view *Shlewis's view*.[2]

Sections 2 and 3 of this paper present the case for Shlewis's view, showing how it can provide attractively unified solutions to similar puzzles about modality and causation. Lewis himself avoids being committed to it only because he doesn't think that modal counterpart theory is the right way to analyse *de re* modal talk about *events*, even if it is the right way to analyse *de re* modal talk about *objects*. Section 4 argues that Lewis's reasons for adopting this disjunctive approach are unconvincing – Shlewis's view is better motivated than Lewis's view, by Lewis's own lights. Finally, in section 5, I argue that Shlewis's view also has a number of important advantages over a closely related

[1] Kim (1976) and Yablo (1992) have defended similar views. Bennett (1988) would argue that the imperfect nominals 'John's saying "Hello"' and 'John's saying "Hello" loudly' denote distinct *facts*, whereas the perfect nominals 'John's saying of "Hello"' and 'John's loud saying of "Hello"' denote one and the same *event*; but for the purposes of this paper I will assume, with the majority, that all causal claims predicate relations of ordered pairs of events (see Chapter One, note 8, above).

[2] For similar approaches to these problems, see McDonnell (forthcoming) and Wasserman (ms).

approach due to Jonathan Schaffer (2005), according to which causal claims are relativized to a pair of sets of 'contrast events'.

The conclusion I am arguing for in this paper is *doubly-conditional*: It's the claim that, *given* a counterfactual account of causation, *if* modal claims are relativized to a choice of counterpart relations, then causal claims are too. I do not endorse, nor will I defend, either of the antecedents of this conditional. But those who do ought to be Shlewisians.

2. Counterpart Theory

Lewis's counterpart theory is best understood within the context of his general project of analysing talk of how things *could* or *must* have been (call this language 'Modalese') in terms of how things are in *possible worlds* (call this language 'Worldese'). Lewis is a realist about possible worlds talk – he thinks that possible worlds *exist*, in a suitably unrestricted sense, that they are spatiotemporally isolated entities of the same kind as our own universe, and that sentences of Worldese (and therefore sentences of Modalese too) are true or false in virtue of how things are in these worlds. The analysis is straightforward for Modalese sentences with no singular terms: To say that there could have been a flying pig is to say that there is a possible world in which there is a flying pig; to say that, necessarily, all bachelors are male is to say that in no possible world is there a non-male bachelor; and so on. What is less straightforward is what to say about those sentences of Modalese which do contain singular terms, sentences like 'George Osborne could have had eleven fingers'.[3] To say that George Osborne could have had eleven fingers is presumably to say of *someone* in some possible world that they have eleven fingers – but whom?

According to Lewis, one shouldn't say that George Osborne could have had eleven fingers if and only if *George Osborne* – the very same! – has eleven fingers in some possible

[3] I'm assuming that proper names are singular terms, of course.

world. Not because there is something incoherent about one thing being a part of many possible worlds – two possible worlds might share George Osborne in the same way that Siamese twins share a hand – but because it would require George Osborne to have exactly ten fingers and exactly eleven fingers, and nothing can instantiate both these properties (Lewis 1986b: 198-209). Nor should one say that George Osborne could have had eleven fingers if and only if George Osborne has a *part* which has eleven fingers. Not because there is something incoherent about 'trans-world' individuals, things that have parts in multiple possible worlds but aren't wholly located in any one of them – given the existence of two objects in different possible worlds, it follows from Lewis's commitment to unrestricted composition that there is something which is their mereological fusion – but because it offends common sense to talk of ordinary objects like George Osborne as scattered across multiple disconnected spacetime regions.[4] Rather, according to Lewis, we should think of ordinary objects as *world-bound* – all their parts are parts of exactly one possible world. There is no possible world in which George Osborne has eleven fingers because George Osborne only exists in the actual world, and in that world he has ten fingers. But objects have *counterparts* in worlds other than their own. What it is to say that George Osborne could have had eleven fingers is to say that there is a counterpart of George Osborne, in some possible world, who has eleven fingers.

A counterpart of an object, according to Lewis, is an object which is *similar* to it, in some respect. Since there are multiple respects in which two objects can resemble each other, it follows that there are multiple counterpart relations. George Osborne, for example, has multiple sets of counterparts. Some of these may contain people with eleven fingers, and some of them may not. Hence the sentence 'George Osborne could have had eleven fingers' may be true relative to one counterpart relation and false relative to another. The counterpart relation relative to which a token modal utterance is evaluated is determined

[4] "[A] modal realism which makes ordinary things out to be trans-world individuals disagrees gratuitously with common opinion." (Lewis 1986b: 220).

by the conversational context, according to Lewis; and *one* of the relevant features of the context is the speaker's choice of object description.[5] Referring to an object one way may 'evoke' one counterpart relation and referring to it a different way may 'evoke' a different counterpart relation.

Lewis exploits the multiplicity of counterpart relations to dissolve a well-known paradox. Imagine a plastic dishpan:

> [T]he plastic is synthesised right in the mold, so it no sooner exists at all than it constitutes the dishpan; and the dishpan is destroyed just when the plastic is incinerated. But suppose the factory had received its order for plastic dishpans a day later. The same bit of plastic would have been made; for the raw materials were already divided into portions just right to fill one mold. But it would have been made in another mold, and it would have constituted a wastebasket. The dishpan would have been made the next day out of different plastic. (Lewis 1986b: 252).

Many would conclude from this example that the dishpan and the dishpan-shaped bit of plastic instantiate different modal properties, and hence that they are distinct objects (Burke (1992) calls this the 'standard account'). But not Lewis:

> It reeks of double counting to say that here we have a dishpan, and we also have a dishpan-shaped bit of plastic that is just where the dishpan is, weighs just what the dishpan weighs (why don't the two together weigh twice as much?), and so on. This multiplication of entities is absurd on its face (Lewis 1986b: 252).

Instead, Lewis claims that he can have it all – he can explain why substituting 'the dishpan' for 'the plastic' in modal contexts doesn't preserve meaning while at the same time insisting that these two descriptions denote the same object. His solution is this:

[5] Other features of the context can play a role too; after all, I can point to a bullet and utter the sentence '*This* could have killed you', and thereby determinately express something true, even though I do not refer to the bullet under any description (Baker 1997: 606).

"We have one thing. What we have two of…are ways of representing" (Lewis 1986b: 249). More precisely:

> We have two candidates to be the counterpart of the plastic which is the dishpan…we have a counterpart relation…whereby the world in question represents the plastic, that is the dishpan, as being made in the form of a wastebasket instead […and] we have another counterpart relation…whereby the world in question represents the dishpan, that is the plastic, as being made a day later out of different raw materials. Here at this world (the only world it is part of) the dishpan just is the plastic; but different ways of referring to this single thing tend to evoke different counterpart relations. (Lewis 1986b: 257).

De re modal claims predicate different properties of an object depending on the counterpart relation relative to which they are evaluated, which *itself* depends on the description under which the object is picked out. Hence, when we utter the sentence 'The plastic could have been a wastebasket' we predicate one property of an object, and when we utter the sentence 'The dishpan couldn't have been a wastebasket' we withhold a *different* property from that very same object. And there's nothing paradoxical about that.

Lewis used the same strategy many times throughout his career in response to many different kinds of puzzles. He appeals to the context-sensitivity of modal predication in explaining how a person can survive the death of their body despite being identical to their body (Lewis 1971) and why Alex is the truthmaker for 'Alex is happy' despite possibly being unhappy (Lewis 2003), for example. But he didn't apply it to the case of John's saying 'Hello'. So I want to describe a view Lewis *might* have held. Since any view Lewis might have held is a view a counterpart of Lewis *does* hold, let's call this counterpart of Lewis 'Shlewis' and let's call this view 'Shlewis's view'.

3. Shlewis's View

According to Shlewis, events, just like objects, are world-bound individuals. They instantiate modal properties in virtue of having counterparts in worlds other than their own. What it is to say that the conference could have been three days long, for example, is to say that the conference has a counterpart, in some possible world, which is three days long.

Shlewis also endorses a counterfactual account of causation. According to the simplest such account, an event *C* is a cause of an event *E* if and only if *E* wouldn't have occurred if *C* hadn't occurred. *Lewis's* official account of causation is much more subtle than that, for a variety of good reasons none of which will be relevant for our purposes (but see Lewis 1986c). For simplicity's sake, therefore, let's suppose that Shlewis endorses the simple definition.

The next step is to analyse counterfactual claims in Worldese terms. Shlewis again subscribes to Lewis's (1973) view on the matter, according to which, granting certain assumptions,[6] if φ and ψ are sentences of Modalese with no singular terms, (φ $\Box\!\!\rightarrow$ ψ) is true if and only if ψ is true in all the closest φ-worlds (the worlds in which φ is true). For example, 'If there had been a flying pig, someone would have seen one by now' is true if and only if someone has seen a flying pig in all the closest possible worlds in which there is a flying pig.

For counterfactuals that *do* include singular terms, things aren't quite so simple. For example, Shlewis can't say that '*C* is a cause of *E*' is true if and only if *E* doesn't occur in all the closest possible worlds in which *C* doesn't occur, since according to Shlewis, *C* and *E* only occur in one possible world, the actual world. What he should say instead, of

[6] In particular, I'm assuming that there is a non-empty set of *closest* φ-worlds, rather than an infinite sequence of closer and closer φ-worlds. Lewis thinks there are good reasons not to make this assumption, and so his official definition is again more subtle than this one. But none of these complications are relevant to my purposes here.

course, is this: '*C* is a cause of *E*' is true if and only if no *counterpart* of *E* occurs in every closest world in which no counterpart of *C* occurs.[7] But if there are multiple counterpart relations, it follows straight away that causal claims are *doubly* relative – they're relativized to a choice of counterpart relation for the cause and a choice of counterpart relation for the effect. A given causal claim may be true relative to one pair of counterpart relations and false relative to another.

Shlewis embraces this conclusion. Moreover, he spots an opportunity to exploit it to solve our causal puzzle. Lewis thinks that the only way of accounting for our causal judgements about the case of John's greeting is to conclude that 'John's saying "Hello"' and 'John's saying "Hello" loudly' denote distinct events. But Shlewis disagrees. It reeks of double counting, he says, to say that here we have a greeting which is essentially loud, and we also have a greeting which is accidentally loud, occurs just where the essentially loud greeting occurs, is *in fact* just as loud (why aren't the two together twice as loud?), and so on. This multiplication of entities is *just* as absurd as the multiplication of entities Lewis complains of in the dishpan case.

Here's what we should say instead, according to Shlewis. 'John's saying "Hello"' and 'John's saying "Hello" loudly' denote the same event. But that event has multiple sets of counterparts. One such set, let's suppose, only contains loud sayings of 'Hello' by John – call the corresponding counterpart relation R_L. Another such set contains only sayings of 'Hello' by John, but not all of them are loud (some of them are quiet) – call the corresponding counterpart relation R_H. Plausibly, on the analysis above, the sentence 'If John's saying "Hello" hadn't occurred, Fred wouldn't have greeted him in return' is true relative to R_H but false relative to R_L. This is because the closest possible worlds in which

[7] We have to be a bit more careful than this if we allow for multiple counterparts *of a single kind* in the same possible world. If the counterpart relation is a similarity relation then it will have this structure, since, for example, there is a possible world containing two distinct individuals which resemble George Osborne in the same way and to the same extent (maybe in this world George Osborne's mother gave birth to identical twins). Lewis (1973: 42) suggests a fix for such cases, but I'll ignore these complications here.

a H-counterpart of John's saying 'Hello' doesn't occur are all ones in which John doesn't say anything, and so ones in which Fred does not greet him in return; whereas the closest possible worlds in which an L-counterpart of John's saying 'Hello' doesn't occur are all ones in which John still says 'Hello', only not so loudly, and so ones in which Fred does greet him in return. Therefore, relative to a suitable choice of counterpart relation for Fred's greeting, the sentence 'John's saying "Hello" caused Fred's greeting' is true relative to R_H but false relative to R_L.[8]

Again, it's plausible to suppose that the conversational context determines the counterpart relations relative to which a causal utterance is evaluated, and that *one* of the relevant features of the context is the speaker's choice of event descriptions. Hence, in some contexts, when we utter the sentence 'John's saying "Hello" caused Fred's greeting' we predicate one relation of an ordered pair of events, and when we utter the sentence 'John's saying "Hello" *loudly* didn't cause Fred's greeting' we withhold a *different* relation from that very same ordered pair. Adding adverbial modifiers to an event description can sometimes change the proposition expressed by a causal utterance in which it is embedded; but if there are multiple counterpart relations, we can explain this without postulating extra events for the modified event descriptions to refer to.[9]

Shlewis's approach is attractively unified. He provides similar solutions to similar problems, using only resources which he takes to be independently motivated. This is

[8] Astute readers will notice that I am playing fast and loose, for the purposes of this paper at least, with the distinction urged in Chapter One between *causing* an effect and being *a cause* of it. If Shlewis subscribes to the position defended in Chapter One, he will think that both modal and causal claims are context-sensitive along *two* distinct dimensions, one having to do with how world-bound entities are represented in non-actual worlds, and the other having to do with restrictions on the set of possibilities over which causal claims are quantifying.

[9] Compare with Kim's strategy for dealing with the differing semantic roles of 'Sebastian's stroll' and 'Sebastian's leisurely stroll' in causal contexts – the "'official line' of the property exemplification account" – which is to affirm that they denote "different, if not entirely distinct, events" (Kim 1976: 319).

very much in the spirit of Lewis's philosophical program (there's a reason why Lewis and Shlewis are counterparts). Yet Lewis did not adopt Shlewis's view. Why not?

4. Lewis vs. Shlewis

Unlike Shlewis, Lewis doesn't think that events are world-bound individuals. Rather, on Lewis's picture, events are *properties of spacetime regions* (Lewis 1986a: 243-7). More carefully: Events are properties, which satisfy the following "conditions of formal eligibility": They are instantiated by "one region each from some worlds, none from others, and never more than one from the same world" (Lewis 1986a: 245). Lewis is clear that this is only intended as a *necessary* condition on eventhood. "Not just any property meeting the conditions given is to count as an event. I have said what kind of things events are – namely, formally eligible properties of regions – but not which things of that kind are events" (Lewis 1986a: 245). To determine how many events there are (that is: to determine how many properties of spacetime regions are also events), Lewis recommends the typically Lewisian strategy of triangulating between intuitions and other metaphysical commitments.

When we try to combine Lewis's metaphysics of events with his other metaphysical commitments, however, problems start to arise. According to Lewis's own analysis, the sentence 'If John's greeting hadn't occurred, Fred's greeting wouldn't have occurred' is true if and only if no counterpart of Fred's greeting occurs in every closest possible world in which no counterpart of John's greeting occurs. But given his nominalism about properties, Lewis's theory of events amounts to the claim that events are *classes* of possible spacetime regions, at most one per world. A class, "like a number…exists necessarily, from the standpoint of every world" (Lewis 1986a: 247). Hence, if what it is for an event to 'occur' in a world is just for it to *exist* in that world, there *is* no possible world in which a counterpart of John's greeting fails to occur, and so it's *vacuously* true that no counterpart of Fred's greeting occurs in every closest possible world in which no

counterpart of John's greeting occurs. Combining Lewis's counterfactual account of causation with his metaphysics of events delivers the absurd conclusion that every causal claim is vacuously true.

To deal with this, Lewis effectively introduces a set of entirely new clauses to his analysis of Modalese, *just* for the special case of Modalese sentences that denote events. To say that an event necessarily occurs, according to Lewis, is not to say that there is a counterpart of the event in every possible world, but rather that the event has an element in every possible world. To say that an event is necessarily a death is not to say that every counterpart of the event is a death but rather that, for every one of the event's elements, "someone dies throughout that region, and not throughout any larger one" (Lewis 1986a: 248). To say that an event necessarily involves Socrates is not to say that every counterpart of the event involves a counterpart of Socrates, but rather that every one of the event's elements contains a counterpart of Socrates. To say that an event necessarily occurs in the library at 12pm is *not* to say that every counterpart of the event occurs in a counterpart of the library at (a counterpart of?) 12pm, but rather that every one of the event's elements is a counterpart of the region occupied by the library at 12pm. "And so on", Lewis (1986a: 248) concludes, although it's by no means obvious that this 'analysis' fully generalises (how, for example, would Lewis analyse the sentence 'Fred's greeting is necessarily an event'?).

Here, in any case, is Lewis's official analysis of 'If John's greeting hadn't occurred, Fred's greeting wouldn't have occurred'. We need to look at the closest possible worlds no parts of which are elements of John's greeting; if no parts of any of those worlds are elements of Fred's greeting, then Fred's greeting wouldn't have occurred if John's greeting hadn't occurred. In other words, John's greeting was a cause of Fred's greeting if and only if: If a world *w* is one of the closest possible worlds no parts of which are elements of John's greeting, no part of *w* is an element of Fred's greeting either. This definition is *not* relativized to a choice of counterpart relations – if a spacetime region is an element of

John's greeting, then it is so *simpliciter*. So although Lewis grants that there is a "persuasive intuition" (Lewis 1986a: 256) that 'John's saying "Hello"' and 'John's saying "Hello" loudly' pick out the same event, he ends up being forced to abandon this intuition in order to account for our causal intuitions about this case.

Shlewis, to repeat, isn't committed to this inelegant, disjunctive analysis of Modalese sentences. He just applies Lewis's own counterpart-theoretic analysis to the special case of Modalese sentences that denote events. So why didn't Lewis do the same? The only hint he provides has to do with his desire to do without events as *sui generis* entities. There is a one-to-one correspondence between events and the classes of spacetime regions in which the events could have occurred. "A one-to-one correspondence is an opportunity for reduction, and I see no reason why events are needed as irreducible elements of being" (Lewis 1986a: 245). Identifying events with properties forces Lewis to accept a disjunctive analysis of Modalese sentences, as well as a multiplication of events that seems just as absurd as the multiplication of objects he is so quick to mock in the dishpan case. This seems to be a price he is willing to pay, however. The question is whether this is a false choice. Is there a similarly reductive metaphysics of events according to which they are *world-bound* individuals that instantiate modal properties in virtue of having counterparts in other worlds? In other words, can Shlewis match Lewis's reductive credentials without abandoning his elegantly unified solution to these modal and causal puzzles?

I think he can. Shlewis could identify events, not with *properties that spacetime regions instantiate*, but rather with *instantiations of properties by spacetime regions*. Call such things 'states of affairs'. This metaphysics of events is very similar to Lewis's. There are exactly as many states of affairs as there are instantiated properties of spacetime regions satisfying Lewis's conditions of eligibility, for example. But states of affairs, unlike properties, are world-bound entities. The spatiotemporal location of a state of affairs is

that spacetime region whose instantiation of a property the state of affairs is identical to, and the world in which a state of affairs occurs is the world in which it is located.

As before, Shlewis should only be interpreted as specifying a *necessary* condition on eventhood. All events are states of affairs, but not every state of affairs is an event. To determine how many events there are (that is: to determine how many states of affairs are also events), Shlewis adopts Lewis's strategy of triangulating between intuitions and other metaphysical commitments. The big difference is that Shlewis has the resources of counterpart theory. Thus he can respect Lewis's own intuition that 'John's saying "Hello"' and 'John's saying "Hello" loudly' denote the same event (that is, the same state of affairs), while also explaining their different semantic roles in causal contexts.

For most of his career, Lewis wouldn't have accepted this proposal, because he didn't believe there were any such things as states of affairs. He agreed that objects instantiate properties, of course; he just denied that, when an object instantiates a property, there is something that is the instantiation of the property by the object. Let *a* be a spacetime region and let F be a property. Suppose, for *reductio*, that there is a state of affairs, S, which is the instantiation of F by *a*. Another central pillar of Lewis's metaphysics is his commitment to the *uniqueness of composition*. All composition is *mereological* composition, for Lewis: "[T]here is only one mode of composition: and it is such that, for given parts, only one whole is composed of them" (Lewis 1986d: 92).[10] But if S is the instantiation of F by *a*, then it must be in some sense 'composed of' F and *a*. Hence it follows from the uniqueness of composition that S = F+*a*, where 'F+*a*' denotes the mereological sum of F and *a*; and there is a "familiar argument" that this is impossible, "else the state of affairs will exist if the thing and the property do, never mind whether the thing instantiates the property" (Lewis 2015: 15). Since mereological composition is unrestricted, *necessarily*, if F and *a* exist, then F+*a* exists. F necessarily exists (it's a class,

[10] It was his finding "unmereological 'composition' profoundly mysterious" that prompted Lewis to analyse even set-theoretic membership in mereological terms (Lewis 1991: 57).

remember), so this amounts to the claim that, necessarily, if *a* exists, then F+*a* exists. But the same isn't true of S. S is necessarily an *instantiation* of F by *a* – it would not have existed if *a* had existed but had not instantiated F. So S and F+*a* instantiate different modal properties; so S ≠ F+*a*; so there are no states of affairs; so events aren't states of affairs.

After repeating this argument in a number of places (Lewis 1999a, 1999b, 2001), however, Lewis eventually retracted it: "I've come around to thinking that a theory of states of affairs…[is] entirely satisfactory" (Lewis 2015: 15). Tellingly, it's counterpart theory to the rescue once again. Suppose the state of affairs 'composed of' F and *a* is just the mereological sum of F and *a*. According to counterpart theory, the modal properties of this thing are relativized to a choice of counterpart relation. "Calling this one thing a sum evokes one counterpart relation…[c]alling the same thing a state of affairs evokes another counterpart relation" (Lewis 2015: 16). Relative to the counterpart relation evoked by calling it a sum, F+*a* is such that it necessarily exists if *a* exists; relative to the counterpart relation evoked by calling it a state of affairs, F+*a* is *not* such that it necessarily exists if *a* exists (it wouldn't have existed if *a* hadn't instantiated F). There is one thing, and two ways of representing.

In summary, Shlewis's treatment of the case of John's greeting is perfectly compatible with Lewis's other metaphysical commitments. Moreover, it has a number of advantages over Lewis's own treatment. It simply applies Lewis counterpart-theoretic analysis of Modalese to the special case of Modalese sentences that denote events, rather than introducing a raft of extra clauses. It respects Lewis's own intuition that 'John's saying "Hello" loudly' and 'John's saying "Hello"' denote the same event, while also accounting for their different semantic roles in causal contexts. And it respects Lewis's own desire to do without events as *sui generis* entities – events, according to Shlewis, are states of affairs, world-bound entities which instantiate modal properties in virtue of having counterparts in worlds other than their own.

To be clear: I don't mean to suggest that Shlewis's view is correct. Nor do I necessarily endorse Lewis's controversial metaphysical methodology, which I've been implicitly assuming on Shlewis's behalf. All I want to point out is that, *by Lewis's own lights*, Shlewis's view is better than the one Lewis actually held. Lewis should have been a Shlewisian.

5. Shlewis vs. Schaffer

Shlewis's approach to the case of John's greeting is similar to a view recently defended by Schaffer (2005). Schaffer is impressed by the ability of 'rather than' locutions to dissolve apparent causal paradoxes. About the case of John's greeting, he'd point out that it seems true to say that John's saying 'Hello' loudly *rather than not at all* caused Fred to greet him in return (rather than not), and it seems false to say that John's saying 'Hello' loudly *rather than quietly* caused Fred to greet him in return (rather than not). Schaffer concludes from this that causation is not a binary relation between events, but "a *quaternary*, *contrastive* relation: *C* rather than **C*** causes *E* rather than **E***, where **C*** and **E*** are nonempty sets of contrast events" (Schaffer 2005: 327).

Although Schaffer doesn't say this explicitly, it's plausible to suppose that the elements of **C*** and **E*** are non-actual and pair-wise incompossible (no two members of **C*** or **E*** occur in the same possible world). Schaffer also doesn't commit himself to any particular semantics for contrastive causal claims, although he does suggest that their truth consists in the right pattern of counterfactual dependence between the events in **C*** and those in **E***. For example, in one account he considers, *C* rather than **C*** causes *E* rather than **E*** if and only if there is a one-one mapping from **C*** to **E*** under counterfactual entailment (Schaffer 2005: 348).

We rarely (if ever) explicitly specify complete sets of contrast events in everyday causal utterances. But Schaffer nevertheless wants to say that utterances of the form '*C* caused *E*' succeed in expressing contrastive propositions, where the sets of contrast events are

supplied in some way by the conversational context. Hence, in some contexts, an utterance of 'John's saying "Hello" loudly caused Fred to greet him in return' expresses the (true) proposition that John's saying 'Hello' loudly rather than not at all caused Fred to greet him in return rather than not, and in others, it expresses the (false) proposition that John's saying 'Hello' loudly rather than quietly caused Fred to greet him in return rather than not.

Shlewis thinks that causal claims are relativized, not to sets of contrasts, but to sets of counterparts. He would therefore tell a different story about the role of 'rather than' locutions – they're just another way of indicating one's intended choice of counterpart relation. For example, in the sentence 'John's saying "Hello" loudly rather than not at all caused Fred to greet him in return', the 'rather than not at all' clause is providing information about the intended counterpart relation: The relation is such that the closest possible world in which no counterpart of John's saying 'Hello' occurs is one in which he doesn't say 'Hello' at all. Shlewis's counterparts and Schaffer's contrasts play very similar semantic roles in their respective theories. But are there any reasons to prefer one theory over the other?

I think there are at least three reasons to prefer Shlewis's view over Schaffer's. The first reason is that Shlewis's view is independently motivated. All it requires is modal counterpart theory as applied to events and a counterfactual account of causation, both of which can be defended independently of Shlewis's treatment of cases like the case of John's greeting. Indeed, Schaffer himself is sympathetic to both these ingredients: In a footnote to the same paper, he claims that "[e]vents are worldbound occurrences. Whether e_1 at w_1 counts as 'the same event' as e_2 at w_2 is a lax and shifty affair" (Schaffer 2005: 355). Shlewis would argue that this very shiftiness can account for all the data the contrastive account is motivated by, without the need for any further metaphysical machinery.

The second reason to prefer Shlewis's view is that it comes with restrictions on the range of allowable interpretations of causal claims that are not present in Schaffer's framework. For example, consider the sentence 'My scratching my nose last year caused my promotion today'. It's very hard to hear a true reading of this sentence (assuming my superiors aren't impressed by nose-scratchings). But the contrastivist thinks there *are* (an infinite number of!) true readings: My scratching my nose last year rather than relocating to Bermuda caused me to be promoted today rather than not, for example, because if I had relocated to Bermuda I wouldn't have been promoted.

Some contrastivists respond to this worry by appealing to conversational salience (e.g. Northcott 2008). While it's true that my scratching my nose last year rather than relocating to Bermuda caused me to be promoted today rather than not, very few utterances of 'My scratching my nose last year caused my promotion today' actually express this proposition, since in practically no conversational context is my relocating to Bermuda a salient contrast event. This is why we find it difficult to accept that 'My scratching my nose last year caused my promotion today' could actually express something true. But this solution doesn't work. The sentence above still sounds false even when the possibility of my relocating to Bermuda *is* raised to conversational salience. Suppose that I was actually considering relocating to Bermuda last year, and I'm now thanking my lucky stars that I didn't – 'My scratching my nose last year caused my promotion today' doesn't seem any less absurd in this context. A more sophisticated story is needed than a mere appeal to conversational salience, and it's not immediately clear what such a story would look like.

This problem is less severe for Shlewis, because not every set of possible events is a set of counterparts. Counterpart relations are *similarity* relations, and not every set of possible objects is a natural resemblance class.[11] We could easily gerrymander a pair of

[11] I'm assuming a 'sparse' conception of similarity here; see Lewis (1983).

counterpart* relations relative to which every closest possible world in which no counterpart* of my scratching my nose occurs is a world in which no counterpart* of my promotion occurs. But these counterpart* relations almost certainly wouldn't be *counterpart* relations, because they almost certainly wouldn't be similarity relations. The events they relate probably wouldn't form anything approaching a natural resemblance class. Hence Shlewis's view can deliver the result that 'My scratching my nose last year caused my promotion today' is false *simpliciter* – that is, false on all interpretations – which seems like the right result. I'm not convinced the contrastivist can make a similar move. That is, I'm not convinced that there are any independently motivated restrictions on what can count as a set of contrast events which would deliver the result that sentences like 'My scratching my nose last year caused my promotion today' are false *simpliciter*.

The final problem for contrastivism, which Schaffer himself acknowledges, is that there is no standard semantic mechanism by which a conversational context can supply *two* sets of contrasts (Schaffer 2013). Usually, contrasts are furnished by something like the *question under discussion* – questions can be thought of as sets of possible answers, which form part of the contextual scoreboard (Groenendijk and Stokhof 1997). But there is usually only one question under discussion in any one context; and in any case, when we're evaluating causal claims, the question under discussion is whether two events are *connected* in particular way, not whether the events themselves occurred. That the events occurred is presupposed in such conversations, and so there is no question under discussion that could generate alternatives to these events.

Shlewis doesn't have to postulate any semantic mechanism for the introduction of contrasts. He doesn't need contrasts. All he needs are different ways of representing world-bound individuals in non-actual worlds. Moreover, he has independent reasons for thinking that the conversational context supplies such things, reasons which Schaffer

himself accepts. So contrastivism isn't just under-motivated – it's also burdened with additional theoretical difficulties that don't apply to Shlewis's view.

6. Conclusion

Lewis's treatment of the dishpan case can be criticised on numerous fronts. One might reject Lewis's counterpart-theoretic approach to *de re* modality, for example (Fara and Williamson 2005; although see Russell 2013). And even if the counterpart-theoretic approach is correct, one might nevertheless object to the whole strategy of exploiting the context-sensitivity of modal talk as a way of avoiding a bloated ontology. For one thing, as Fine (2003) points out, substituting 'the dishpan' for 'the plastic' apparently fails to preserve meaning in non-modal contexts as well as modal contexts. Consider the following pairs of sentences, for example:

(1a) The dishpan is useful.
(1b) The lump of plastic is useful.

(2a) The dishpan is well-made.
(2b) The lump of plastic is well-made.

Arguably, (1a) and (2a) are true whereas (1b) and (2b) are false.[12] It doesn't immediately follow from this that 'the plastic' and 'the dishpan' pick out different objects, of course, but counterpart theory is clearly of no help here. This invites the accusation that Lewis's solution in the modal case is just *ad hoc*.

For another thing, as Hawthorne points out, one cannot generally evaluate a single instance of a context-sensitive term in a sentence with respect to two different contextual parameters. "'The new power forward for the Denver Nuggets is not tall but my best friend is' cannot be used felicitously in a case where my best friend is also the new power

[12] Personally, my intuitions about (1) and (2) are less clear than my intuitions about the corresponding modal claims, but I take it that this judgement isn't universally shared.

forward and is tall for a person but not tall for a basketball player", for example; nor can I utter "'My best friend but not the lord mayor is going to a nearby bar' in a case where 'my best friend' and 'the lord mayor' pick out the same object" (Hawthorne 2008: 386). By contrast, 'The plastic, but not the dishpan, could have been a wastebasket' sounds just fine to me, even though, according to Lewis, its truth would require one token instance of a modal predicate to be evaluated with respect to two different counterpart relations. Even more obviously, consider sentences like the following: 'It could have been the case that the plastic was a wastebasket but the dishpan wasn't'. This sentence certainly *seems* to be saying *of a conjunction* – that the plastic is a wastebasket and the dishpan isn't – that it is possibly true. But if the dishpan is identical to the plastic, this conjunction is of course necessarily false, since it's saying of a single object that it both is and isn't a wastebasket. It's not at all obvious that the sentence above is necessarily false, however.

Shlewis's treatment of the case of John's greeting can be criticised on exactly the same grounds. Substituting 'John's saying "Hello"' with 'John's saying "Hello" loudly' seems to fail to preserve meaning in non-causal contexts as well as causal contexts – there are ways of filling in the details of the case so that (3a) seems true whereas (3b) seems false, for example:

(3a) John's saying "Hello" loudly was rude.

(3b) John's saying "Hello" was rude.

Moreover, the sentence 'It was John's saying "Hello", and not his saying "Hello" loudly, that caused Fred to greet him in return' sounds just fine, even though its truth, according to Shlewis, requires a single instance of 'caused' to be evaluated simultaneously with respect to two different counterpart relations.

Recall my doubly-conditional conclusion: *Given* a counterfactual account of causation, *if* modal claims are relativized to a choice of counterpart relations, then causal claims are too. Let me emphasise again that I have provided no reasons for believing either of the

antecedents of this conditional. They might well be false. And even if they're true, it doesn't follow that the strategy of exploiting the context-sensitivity of modal and causal talk as a way of avoiding a bloated ontology is correct. But there's a right way to carry out this strategy, and it's Shlewis's way. There are some philosophers who either do or should endorse the antecedents of my conclusion and yet fail to appreciate the force of its consequent. There's a general moral to this story, even for those with little sympathy for Shlewis's view: Some puzzles about causation can be solved simply by paying close attention to the relationship between causation and modality.

References

Baker, L. R. (1997). Why constitution is not identity. *Journal of Philosophy 94*(12), 599-621.

Bennett, J. (1988). *Events and their names*. Indianapolis, IN: Hackett Publishing.

Burke, M. (1992). Copper statues and pieces of copper: A challenge to the standard account. *Analysis 52*(1), 12-17.

Fara, M. and Williamson, T. (2005). Counterparts and actuality. *Mind 114*(453), 1-30.

Fine, K. (2003). The non-identity of a material thing and its matter. *Mind 112*(446), 195-234.

Groenendijk, J. and M. Stokhof (1997). Questions. In J. van Bentham and A. ter Meulen (eds.), *Handbook of logic and language*, Amsterdam: Elsevier.

Hawthorne, J. (2008). Three-dimensionalism vs. four-dimensionalism. In T. Sider, J. Hawthorne and D. W. Zimmerman (eds.), *Contemporary debates in metaphysics*, Oxford: Blackwell.

Kim, J. (1976). Events as property exemplifications. In M. Brand and D. Walton (eds.), *Action theory*, Dordrecht: D. Reidel.

Lewis, D. (1971). Counterparts of persons and their bodies. *The Journal of Philosophy 68*(7), 203-211.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1976). Survival and identity. In A. O. Rorty (ed.), *The identities of persons,* London: University of California Press.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy 61*(4), 343-377.

Lewis, D. (1986a). Events. In D. Lewis, *Philosophical papers (vol. 2)*, Oxford: Oxford University Press.

Lewis, D. (1986b). *On the Plurality of Worlds*. Oxford: Blackwell.

Lewis, D. (1986c). Causation. In D. Lewis (ed.), *Philosophical papers (vol. 2)*, Oxford: Oxford University Press.

Lewis, D. (1986d). A comment on Armstrong and Forrest, *Australasian Journal of Philosophy 64*(1), 92-93.

Lewis, D. (1991). *Parts of classes.* Oxford: Basil Blackwell.

Lewis, D. (1999a). Armstrong on combinatorial possibility. In D. Lewis, *Papers in metaphysics and epistemology*, Cambridge: Cambridge University Press.

Lewis, D. (1999b). A world of truthmakers?. In D. Lewis, *Papers in metaphysics and epistemology*, Cambridge: Cambridge University Press.

Lewis, D. (2001). Truthmaking and difference-making, *Noûs 35*(4), 602-615.

Lewis, D. (2003). Things qua truthmakers. In H. Lillehammer and G. Rodriguez-Pereyra (eds.), *Real metaphysics: Essays in honor of D. H. Mellor*, Abingdon: Routledge.

Lewis, D. (2015). Counterparts of states of affairs. In B. Loewer and J. Schaffer, *A companion to David Lewis*, Oxford: Blackwell.

McDonnell, N. (forthcoming). Events and their counterparts. *Philosophical Studies*.

Northcott, R. (2008). Causation and contrast classes. *Philosophical Studies 139*(1), 111-123.

Russell, J. S. (2013). Actuality for counterpart theorists. *Mind 122*(485), 85-134.

Schaffer, J. (2005). Contrastive causation. *Philosophical Review 114*(3), 327-358.

Schaffer, J. (2013). Causal contextualism. In M. Blaauw (ed.), *Contrastivism in philosophy*, New York: Routledge.

Wasserman, D. (manuscript). Is causation extensional?.

Yablo, S. (1992). Cause and essence. *Synthese 93*(3), 403-449.

Chapter Four

Interventionism and Mental Surgery

Abstract: Campbell (2007) has recently argued that the interventionist account of causation must be amended if it is to be applied to causation in psychology. The problem, he argues, is that it follows from one of the conditions on interventions – the so-called 'surgical' constraint – that an intervention on an intention requires the suspension of the agent's rational autonomy. In this paper, I argue that the problem Campbell identifies is in fact an instance of a wider problem for interventionism, extending beyond psychology, which I call the problem of 'abrupt transitions'. I then defend a solution to the problem, which replaces the surgical constraint with a weaker constraint on interventions that nevertheless does all the work the surgical constraint was designed to do. I conclude by exploring some interesting consequences of this weaker constraint for causation in psychology.

1. Introduction

The interventionist account of causation is an analysis of causal claims in terms of correlations under *interventions*, manipulations of variables satisfying certain explicitly causal conditions. Campbell (2007) has recently argued that one of these conditions – the so-called 'surgical constraint' – runs into trouble when we try to apply interventionism to causation in psychology. The problem, he argues, is that the surgical constraint requires that an intervention on an intention should remove that intention from the influence of its rational causes; and it's implausible that our interest in psychological causation is an interest in what would happen in situations in which the rational autonomy of an agent is suspended in this way.

This paper aims to do two things. First, I will argue that the problem Campbell raises for interventionism is in fact an instance of a wider problem, extending beyond psychology, which I call the problem of 'abrupt transitions' after a similar problem discussed by David Lewis. Second, I'll argue that the problem of abrupt transitions can be solved by replacing the surgical constraint with a weaker constraint on interventions, one which nevertheless does all the work the surgical constraint was designed to do. I conclude by exploring some interesting consequences of this weaker constraint for causation in psychology.

2. Interventionism Introduced

Barometer readings are correlated with weather patterns. When barometer readings go down, stormy conditions tend to follow; and when barometer readings go up, fair weather conditions tend to follow. That, after all, is why we use barometers to forecast the weather. But correlation does not imply causation, to use a well-worn cliché. A fall in the reading of a barometer is not a cause of the storm which follows it.

According to the simple counterfactual account of causation, the fall in the barometer reading was a cause of the storm if and only if the storm wouldn't have occurred but for the fall in the barometer reading; and according to Lewis (1973), *this* counterfactual is true if and only if the storm doesn't occur in all the closest (accessible) possible worlds in which the fall in the barometer reading doesn't occur. In selecting the closest possible worlds, Lewis (1979: 472) suggests that "[i]t is of first importance to avoid big, widespread, diverse violations of law…[and it] is of second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails". The closest possible worlds in which the barometer reading doesn't fall, for example, are ones which are just like the actual world until a small 'miracle'[1] occurs in the barometer a few seconds earlier, which prevents the needle from moving – these are

[1] See Chapter Two, Section 3, above, for more on what Lewis means by 'miracle' here.

the worlds that best satisfy the dual priorities of minimizing violations of the actual laws and minimizing deviations from actual matters of fact. Since the storm still occurs in at least one of these worlds, it follows that the fall in the barometer reading wasn't a cause of the storm, as required.

As Edgington (2011) points out, however, Lewis's analysis doesn't fare so well in other cases. Imagine that I am measuring, by means of the sturdiest instrument in the world, the pressure of a tiny amount of extremely volatile gas. Suppose there's a change in both the pressure of the gas and the reading of my instrument. What is the closest possible world in which the instrument reading changes? One candidate is the world in which a small miracle occurs in the instrument. But this world is arguably trumped on Lewis's closeness metric by the world in which a small miracle occurs in the *gas* a few seconds earlier, which then causes a change in the instrument reading. The first world does slightly better than the second in terms of maximizing match with actual matters of fact, but the second world does *much* better in terms of minimizing violations of the actual laws, since the instrument, being sturdy, needs a 'large' miracle to change its reading, whereas the gas, being volatile, only needs a 'small' miracle to change its pressure. Lewis's analysis therefore falsely implies that the change in the instrument reading was a cause of the earlier change in the pressure of the gas.[2]

Let's take a step back. How do we actually *know* that barometer readings don't cause storms? Ask a scientist, and they'll tell you that barometer readings and weather patterns are correlated only because they have a *common* cause, namely, the atmospheric pressure; and moreover, that if were we to change the reading of a barometer in such a way that doesn't also change the atmospheric pressure – by opening up the barometer and fiddling about with the needle, for example – it would

[2] Lewis could plausibly respond to this objection either by tweaking his definition of closeness or by saying more about what it is for a miracle to be 'large' or 'small', but it's not at all clear that there is a systematic solution available to these kinds of problems.

make precisely no difference to the weather. It's by means of exactly these kinds of targeted manipulations,[3] informed by prior causal knowledge, that causal hypotheses are actually tested in the natural and social sciences.

Although every account of causation will, of course, find some connection between causal relationships and correlations under experimentally idealized manipulations, interventionism is unique in taking such correlations to be *constitutive* of causation, and not simply evidence for it.[4] Roughly speaking, the interventionist account defines causation, not in terms of what would have happened if the cause hadn't *occurred*, but rather in terms of what would have happened if the cause had been prevented from occurring *by means of an intervention*. The concept of an intervention is then itself defined in causal terms. As an analysis of causal claims, then, interventionism is explicitly and unashamedly non-reductive. But it nevertheless purports to be illuminating – in just the sense that functionalism about the mind seeks to better understand mental states in terms of the functional relations between them and not simply by defining each one individually in behavioural terms, interventionism seeks to better understand both causation and experimental manipulation in terms of the connections between them and not simply by defining each of them individually in non-causal terms.[5, 6]

---

[3] Or, more commonly, by means of statistical techniques that allow us to *simulate* the effects of these kinds of targeted manipulations.

[4] One can draw a useful analogy with frequentism about probability here, which similarly attempts to treat long-run relative frequencies as constitutive of objective probability, and not simply evidence for it – see Wallace (2012: ch.4) for helpful discussion.

[5] C.f. (Woodward 2003: 106): "Despite the failure of psychological claims to reduce to claims about overt behaviour or probability claims to reduce to claims about relative frequencies, we can get a nontrivial purchase on what such claims mean by understanding how, given appropriate additional assumptions, they connect with facts about behaviour or facts about frequencies."

[6] This analogy suggests an interpretation of interventionism as a first step in a fully-fledged reductive analysis of causation along the lines of the 'Canberra Plan' (see Lewis 1970; Jackson 1998; the term was first coined by O'Leary-Hawthorne and Price 1996). The idea is that we would take the conjunction of all our interventionist definitions, 'Ramsify out' all the causal predicates by replacing them with a higher-order variables bound by a higher-order existential quantifiers, and then search for non-causal properties that witness the resulting quantificational sentence. Although interventionism "might be supplemented by any one of a

A few minor differences notwithstanding, my statement of interventionism will closely follow Woodward's (2003). Woodward's approach is to combine elements from two distinct traditions – the philosophical tradition of analysing causation in terms of manipulability,[7] and the use of *causal models* as vehicles for representing causal structures in econometrics, computer science and experimental design.[8] A causal model is an ordered pair $\mathfrak{M} = \langle \mathbf{V}, \mathbf{E} \rangle$, where **V** is a set of variables and **E** is a set of 'structural equations', one for every variable in **V**. The variables represent "properties or magnitudes that, as the name implies, are capable of taking more than one value" (Woodward 2003: 39). The simplest kind of variable is one which can take two values – say, 0 and 1 – such that $X=1$ if a particular event occurs and $X=0$ otherwise. But multi-valued variables are possible too. For example, we might represent the causal structure of a car by constructing a model which includes a variable X that can take any positive real number as a value, such that $X=x$ if and only if the mass of the car is $x$kg.

The structural equations describe, for each variable $V \in \mathbf{V}$ and each combination of values of the other variables in **V**, what the value of V *would* be if the other variables *were* assigned that combination of values by means of interventions with respect to V. (I haven't said what an intervention is yet – hold tight, we'll get there!) Here's an example of a structural equation:

$$X := Y + Z$$

number of different stories about metaphysical foundations", however, it "does not attempt to provide such foundations", according to Woodward (2008: 194-5).

[7] See Gasking (1955); Collingwood (1940); von Wright (1971); Menzies and Price (1993). Critics of these views (e.g. Hausman 1986) tend to describe them as overly anthropocentric in tying causation too closely to human agency. Woodward's account attempts to meet this objection by defining the notion of an intervention in causal, rather than agential, terms.

[8] See especially, Pearl (2000); Spirtes, Glymour and Scheines (2000).

(1) says that, for every combination of values of Y and Z, were they to be assigned those values by means of interventions with respect to X, the value of X would be equal to the sum of Y and Z.

Relative to a causal model, Woodward starts by defining the notion of a 'direct cause':[9]

> DIRECT CAUSATION (DC): X is a direct cause of Y in $\mathfrak{M} = \langle \mathbf{V}, \mathbf{E} \rangle$ if and only if there is a possible intervention on X with respect to Y, and some combination of values of the other variables in **V**, such that according to **E**, were the intervention to occur while all of the other variables in **V** were held fixed at that combination of values by interventions with respect to Y, there would be a change in the value of Y.

Why the insistence on holding other variables fixed? Because of so-called 'failures of faithfulness' (Spirtes *et al.* 2000). Here's an example due to Hesslow (1976). Whether or not a person is pregnant (P) is known to affect the level of blood-clotting proteins produced by the body, which in turn affects the risk of developing deep vein thrombosis (T). Whether a patient is taking oestrogen-based contraceptive pills (C) is also known to affect the production of blood-clotting proteins. But suppose the negative effect of using contraceptive pills on the chance of developing thrombosis along the $C \rightarrow P \rightarrow T$ route is *exactly cancelled out* by the positive effect of using contraceptive pills on the chance of developing thrombosis along the direct $C \rightarrow T$ route, so that the net effect of changing C on the value of T is zero. Nevertheless, (DC) delivers the correct result that C is a direct cause of T in the model containing these three variables, because were we to intervene on whether or not the patient is using contraceptive pills *while holding fixed* whether the patient becomes pregnant – by replacing the use of contraceptive pills with a different form of contraception, for example – there *would* be a change in the chance of the patient developing thrombosis.

[9] See Woodward (2003: 55). Here and elsewhere, I diverge slightly from Woodward's exact wording of the definitions for ease of exposition.

We can represent causal models by means of *causal graphs*. A causal graph contains a node for every variable of the model and a directed edge between two nodes for every direct causal relationship. Here, for example, is the causal graph of the model containing *C*, *P* and *T* in the example above:

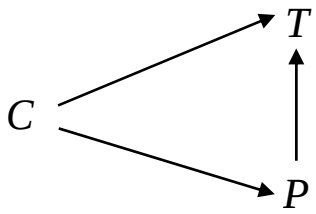

Fig. 1

The concept of direct causation can be used to define the notion of a *path* between variables. An ordered n-tuple of variables $<V_1, V_2,..., V_n>$ is a path from V_1 to V_n in $\mathfrak{M}$ if and only if V_1 is a direct cause of V_2 in $\mathfrak{M}$, V_2 is a direct cause of V_3 in $\mathfrak{M}$, and so on, up to V_n. In the causal model illustrated in fig.1, for example, there are two paths from *C* to *T*, one of which goes through *P*.

We can then define a more general relation between variables as follows:

VARIABLE-LEVEL CAUSATION (VC):[10] *X* is a variable-level cause of *Y* in $\mathfrak{M}$ if and only if:

– there is a path P from *X* to *Y* in $\mathfrak{M}$, and

– there is a possible intervention on *X* with respect to *Y*, and some combination of values of the other variables in **V** not on P, such that according to **E**, were the intervention to occur while all of the other variables in **V** not on P were held fixed at that combination of values by interventions with respect to *Y*, there would be a change in the value of *Y*.

For example, in the model illustrated in fig.2 below, *X* is a variable-level cause of *Y* if and only if there is an intervention on *X* with respect to *Y* which results in a change in the value of *Y* when the value of *Z* (but not, of course, the value of *W*) is held fixed at some value.

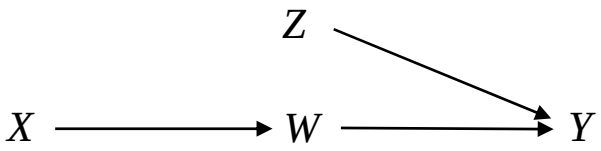

Fig. 2

Having defined a relation between variables, Woodward then defines a relation between *actual values* of variables, as follows (Woodward 2003: 77):

[10] Woodward (2003: 57) calls this relation "type-level causation", but this is a misnomer. Type-causal claims, at least as the term is generally used in the philosophy literature, are claims like 'Smoking causes cancer'. These are *generic* claims, of the same semantic class as 'Chickens lay eggs' or 'Lying is wrong'. The characteristic feature of generics is that they admit of exceptions – 'Chickens lay eggs' is plausibly true, for example, even though fewer than half of all chickens lay eggs. Similarly, it's true that smoking causes cancer, even though not every token smoking event causes a token cancer event (for an introduction to the semantics of generics, see Leslie (2012)). Variable-level causal claims, however, are claims like 'The number of cigarettes smoked by John is a cause of his chance of developing cancer'. These are *not* generic claims –rather, they predicate a relation between two token variables. See also Menzies (2008: 206) on this point.

ACTUAL CAUSATION (AC): $X=x$ is an actual cause of $Y=y$ in $\mathfrak{M}$ if and only if:

- the actual values of X and Y are x and y, respectively,
- there is a path P from X to Y in $\mathfrak{M}$, and
- there is a possible intervention on X with respect to Y such that according to **E**, were the intervention to occur while all of the other variables in $\mathfrak{M}$ not on P were held fixed at their *actual* values by interventions with respect to Y, there would be a change in the value of Y.

You'll notice that (VC) and (AC) are both relativized to a causal model. There have been some attempts to 'de-relativize' these definitions to arrive at a definition of causation *simpliciter* (see Weslake forthcoming; Woodward 2008), a project which is likely to require some kind of restriction to 'apt' or 'appropriate' causal models (see Hitchcock (2001); some are happy to conclude that causation is a fundamentally model-relative concept – see Halpern and Pearl (2005: 845), for example). This is unfinished business for interventionism, but it won't be relevant here.

3. Surgical Interventions

Informally, Woodward describes an intervention on X with respect to Y as a process that satisfies "whatever conditions must be met in an ideal experiment designed to determine whether X causes Y" (Woodward 2003: 46). Formally, it is a change in the value of an *intervention variable for X with respect to Y* from its 'off' value to its 'on' value. Let's say that a model $\mathfrak{M} = \langle \mathbf{V}, \mathbf{E} \rangle$ is 'veridical' if and only if **E** correctly describes the counterfactual relations between the variables in **V**. Then a variable I is an intervention variable for X with respect to Y if and only if:[11]

[11] See Woodward (2003: 98). Woodward drops the reference to models in his statement of these conditions; I follow Weslake (forthcoming) in correcting this omission.

In every veridical model $\mathfrak{M}$ containing I, X and Y:

(*I1*) I is a variable-level cause of X in $\mathfrak{M}$;

(*I2*) When I takes its 'on' value, the value of I is the only actual cause of the value of X in $\mathfrak{M}$;

(*I3*) Every path from I to Y in $\mathfrak{M}$ (if there is one at all) goes through X;[12]

(*I4*) I is statistically independent of every variable in $\mathfrak{M}$ that is on a path to Y that does not go through X.

To motivate these conditions on interventions, consider the following example. It's well-established (e.g. Coleman and Hoffer 1987) that a high school student in the USA will on average receive better grades on standardised tests if she attends a private school than if she attends a state school. But whether this is a *causal* correlation is a hotly debated question in social science. Suppose we tried to test whether private school attendance is a cause of better educational outcomes by means of the following experiment. We recruit a number of participants with children approaching high school age, who cannot afford to send their children to private schools. We then divide these participants into two groups. To the participants in one group, we give a large amount of money (enough to cover private school fees). We then compare the final grades of the children in the first group with the final grades of those in the second.

This, of course, would be a terrible experiment, for a number of reasons. First, we can't be sure that increasing parental income isn't itself an *independent* cause of better grades, for reasons that have nothing to do with private schools. Perhaps, for example, increasing parental income also has the effect of relieving pressure on children to supplement household income through part-time work, freeing up more time to study. So if we *do* observe a difference in final grades between the children in the two groups,

[12] The parenthetical is important here – it's important to note that none of the conditions on interventions imply that I is an intervention variable for X with respect to Y only if X is a cause of Y. If they did, (VC) would be viciously circular – the *definiendum* would be contained in the *definiens*.

we can't be sure that this is because of a causal connection between private school attendance and better educational outcomes, as opposed to an independent causal connection between parental income and educational outcomes.

(*I3*) is designed to address this problem. What (*I3*) requires is that an intervention variable for *X* with respect to *Y* should not cause *Y* 'directly', but only, if at all, 'through' *X*. Let *I*=1 if the parent of a particular child is given the large sum of money and 0 otherwise, let *S*=1 if the child attends private school and 0 otherwise, let *G* be a variable representing the child's final grades, and let *W*=1 if the child is in part-time work and 0 otherwise. Here is the graph of the (veridical) model that includes these four variables (the question mark indicates the causal relation being analysed):

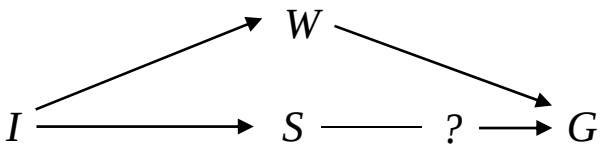

Fig. 3

Since there is a path from *I* to *G* in this model that does not go through *S*, it follows from (*I3*) that *I* is not an intervention variable for *S* with respect to *G*. Variables like *W* that are on paths to the purported effect variable that don't also go through the target variable are sometimes called 'confounders' in the experimental literature, and the need to ensure that our intervention variable is not a cause of such variables is sometimes referred to as the need to 'control for confounders'.

(*I4*) is designed to address a similar problem. Suppose the money for the experiment was raised by cutting funding for state schools. Let *C* be a variable such that *C*=1 if the cuts occur and *C*=0 otherwise. Since such cuts are likely to have an impact on the difference between private and state school grades, the causal model including this variable plausibly looks like this:

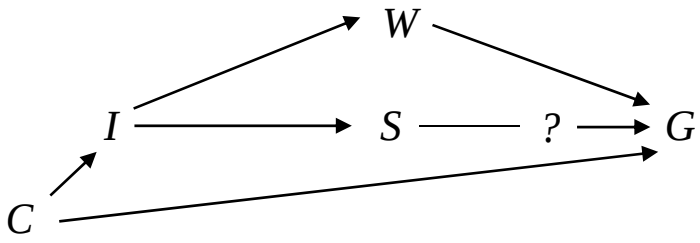

Fig. 4

(*I4*) rules out such scenarios. It requires that an intervention variable *I* for *S* with respect to *G* should be statistically independent of *C* – and one way *I* can fail to be statistically independent of *C* is for *C* to be a cause of *I*.

There is one final problem with my fictional experiment: We can't be sure that the participants in the first group will actually *use* the money we give them to send their children to private school. There might be other reasons – political objections to private education, for example – why these parents might decide to send their children to state schools after all, despite now having the resources to send them to private schools. Hence if we *don't* observe a difference in final grades between the two groups, we can't be sure that this is because of a *lack* of causal connection between private school attendance and better educational outcomes, as opposed to a failure of the intervention to have any significant effect on private school attendance.

(*I2*) is designed to address this problem. It's sometimes described as the requirement that an intervention on *X* be 'surgical', in Pearl's (2000) words. An intervention should amount to "lifting *X* from the influence of the old functional mechanism" in which it was embedded "and placing it under the influence of a new mechanism" (Pearl 2000: 70), one which "breaks whatever endogenous casual relationships are at work" (Woodward 2003: 135) in determining the value of *X*. In other words, an intervention on *X* should ensure that "*X* ceases to depend on the values of the other variables that cause *X*" (Woodward 2003: 98), so that its value "is determined completely by our

intervention, the causal influence of the other variables being completely overridden" (Hitchcock 2012).

If *P* is a variable representing the political commitments of the parents, the graph of the model containing *I*, *P*, *S* and *G* looks like this:

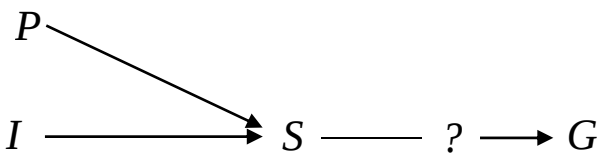

Fig. 5

According to (*I2*), *I* is an intervention variable for *S* with respect to *G* only if, when *I* takes its 'on' value, the value of *I* is the only actual cause of the value of *S* in this model. In particular, when *I*=1, the actual value of *P* should *not* also be a cause of the actual value of *S* – changing the actual value of *P* should have no effect on the actual value of *S*. An intervention on whether a student attends private school, in other words, should ensure that the student attends private school *regardless* of the political commitments of her parents. Simply giving the participants a large sum of money fails to achieve this, because whether or not a child attends private school might continue to depend on the parents' political commitments even after the money is handed over.

4. Interventionism and Rational Causation

It's (*I2*) – the so-called 'surgical constraint' – that Campbell thinks leads to trouble for causation in psychology. Suppose we're interested in whether S's *intention* to φ was a cause of her φ-ing. According to (AC), this is the case only if, had an intervention on whether or not S intends to φ been performed, there would have been a change in whether or not S φ-s. But what is it to intervene on whether someone has the intention to φ?

> We would naturally think of this in terms of providing someone with reasons to φ, or reasons not to φ. 'You think φ-ing will make you happy, but it won't', you might say as an opening move. And you might present further considerations in favor of your remark. You would be appealing to the rationality of the subject. The trouble with this is that it leaves intact the factors that are the usual causes of the someone's forming, or not forming, the intention to do something.
>
> For example, suppose that one of the usual causes of a person's intending to φ is that they think φ-ing will make them happy. If your 'intervention' takes the form of arguing about whether or not φ-ing will in fact make that person happy, then you have left in place one variable that is a usual cause of whether the person forms the intention to φ. This means that the intervention is not, in Pearl's terms, 'surgical'…[Hence,] if an endogenous cause of whether someone forms the intention to φ is whether the person believes that φ-ing will make them happy, then a manipulation of whether the person forms the intention that proceeds by manipulating whether the person believes that φ-ing will make them happy does not constitute an intervention (Campbell 2007: 61-2; with minor changes in notation).

Simply telling S that φ-ing will make her happy doesn't count as an intervention on whether S intends to φ, because it fails to satisfy (*I2*) – whether S intends to φ continues to depend on whether S believes that φ-ing will make her happy, even once the intervention is performed. So what, then, would an intervention on whether S intends to φ actually look like?

> [Such an] intervention would have to come from outside and seize control of whether the subject had the intention, suspending the influence of the subject's usual reasons for forming an intention, such as whether the subject had reasons for forming the intention to φ…This is evidently quite an unusual situation. It does not happen very often, if it happens at all, that a person's rational autonomy is suspended and some alien force seizes control over whether that person has a particular intention. (Campbell 2007: 62; with minor changes in notation).

There are some philosophers who would deny that such an intervention on intentions is even possible. They would deny, in other words, that it's even coherent to speak of intentions to φ being manipulated independently of preferences or beliefs about the consequences of φ-ing. On this form of 'mental holism', the content of an intention is partly *constituted* by the position of that intention in a broadly rational 'web' of mental states. As Campbell (2010: 71) puts it, "[t]he mind has to be organized in a broadly rational way, for there to be a mind there at all". There is no such thing, nor could there be such a thing, as an agent who believes that φ-ing will make her happy, has no reason not to φ, yet nevertheless fails to intend to φ, because it is an *a priori* prerequisite on something exemplifying mental states at all that their mental states fit together in a broadly rational way. In a slogan: One cannot simply pick-and-mix mental states. Therefore, a manipulation of whether or not an agent intends to φ which satisfies (*I2*) is conceptually impossible. And therefore, at least on orthodox semantics of counterfactuals, the counterfactual delivered by the interventionist analysis of, say, 'Jane's intention to dance caused her to dance' is vacuously true.

As Campbell remarks, it's this kind of view of the mental that "underpins some of the hesitation philosophers have felt in talking about mental causation at all" (Campbell 2007: 63). But Campbell himself is no friend of mental holism (see Campbell 2010). He grants that a surgical intervention on an intention to φ is *coherent*. His point is rather to emphasise just how strange such a thing would have to be. What we are asking for is a manipulation of an agent's intentions that disrupts the sensitivity of her intentions to rational evaluation. Indeed:

> Someone who seemed to find him- or herself in that situation – someone who encountered in introspection an intention that seemed to have been the direct result of someone else's long-standing objectives, interests, preferences, and so on – would experience this as *thought insertion*, the feeling that someone else's token thought has been pushed into your mind, one of the symptoms of schizophrenia…It is exactly this situation that we are

> envisaging, though, when we think in terms of surgical intervention on possession of an intention. (Campbell 2007: 62).

Campbell concludes that "it is not credible that our interest in psychological causation is an interest in what would happen under such idealized conditions of alien control" (Campbell 2007: 62). Although he is happy to grant that the interventionist account is *extensionally adequate* (in that it doesn't misidentify a cause as a non-cause or *vice versa*), Campbell nevertheless considers it implausible to suppose that when we talk of Jane's intention to dance as causing her dancing, we're saying something about what would have happened in cases so far removed from our own psychological lives as to be virtually unrecognisable.

In section 6, I will describe what I think is the right solution to Campbell's problem. But first, I want to get clearer on what exactly the problem is. In particular, I want to argue that – contrary to what Campbell seems to suggest – the problem is not confined to causation in psychology.

5. The Problem of Abrupt Transitions

Consider the following example.[13] Tushar is driving in the outside lane of a two-lane road and realizes too late, at time *t*, that he needs to take the next exit. He misses the exit, and as such he is late for his meeting. Tushar's being in the outside lane at *t* was a cause of his lateness. According to (AC), this is so because had an intervention on which lane he was in at *t* been performed, there would have been a change in whether or not he was late for his meeting.

So what would it be to intervene on which lane Tushar was in at *t*? One natural thought is something like this: We call him up some time before *t* and remind him that he needs to take the next exit. But it seems this manipulation wouldn't satisfy (*I2*). Whether

[13] This case is taken from Woodward (2003: 142-4), but was apparently originally discussed by David Lewis in an unpublished lecture.

Tushar was in the outside or the inside lane at t-1, after all, is a cause of whether he was in the outside or the inside lane at t in some model, for any sufficiently fine-grained units of time. An intervention on which lane Tushar was in at t should therefore remove this variable from the influence of his location at any earlier time. In other words, the intervention must ensure that Tushar ends up in the inside lane at t, *wherever* he happens to be at any arbitrarily small time before that. One can certainly *imagine* interventions like that – we could place some kind of portal in the outside lane, for example, which would have the effect of instantly teleporting Tushar into the inside lane if he happens to go through it. (We'd probably also have to do some other things to ensure that this intervention doesn't have any independent effect on whether or not Tushar is late for his meeting, in violation of (*I3*) – for example, we'd probably have to wipe the memories of the drivers around Tushar, so that they don't become too alarmed at the sudden disappearance and reappearance of his car and crash into him, thereby making him late for his meeting.) But this would be a strange intervention indeed. Even if it's *coherent* to imagine a manipulation of Tushar's position at t which removes it from under the influence of his position at earlier times, and even if the interventionist account gets the right result in this case (that Tushar's being in the outside lane at t was a cause of his lateness), it still seems odd that when I talk of Tushar's being in the outside lane at t as a cause of his lateness, I'm saying something about what would have happened in such scenarios, involving portals and mind-wiping devices, so far removed from our everyday experience as to be virtually unrecognisable.

The point here seems quite general. An intervention on whether or not Luis Suarez scores a goal, for example, should ensure that he scores regardless of whether he is substituted a minute beforehand, according to (*I2*); an intervention on whether or not a tree grows in my garden should ensure that it grows regardless of the presence or otherwise of a seed in the soil (or indeed the soil itself); and so on. If (*I2*) is applied at

a sufficiently fine-grained level of detail, interventions start to look *really* weird. Call this the *problem of abrupt transitions*. I submit that the problem Campbell identifies for interventionism in psychology – *at least* if we are unmoved by mental holist worries about the independent manipulability of mental states – is just a special case of this problem.

It's notable that Lewis himself seemed to feel the force of the problem of abrupt transitions. When constructing his closeness metric on possible worlds, Lewis argued that "we should sacrifice the independence of the immediate past to provide an orderly transition from actual past to counterfactual present and future" (Lewis 1979: 463).[14] To determine whether Tushar's being in the outside lane at *t* caused him to be late for his meeting, for example, we should consider a possible world which starts diverging from the actual world a few seconds before *t*, and then smoothly transitions into a world in which he is in the inside lane at *t*. The interventionist account, however, seems inconsistent with such 'transition periods' between actual past and counterfactual future. Indeed, Woodward explicitly acknowledges that, "in contrast to Lewis, the interventionist account tells us that we should avoid transition periods entirely" (Woodward 2003: 144).

6. Indirect Interventions

Campbell's solution to the problem of abrupt transitions is to abandon the surgical constraint on interventions entirely (at least in the special case of psychological causation). This, I think, is an overreaction. Manipulations of variables which satisfy all the conditions on interventions *except* (*I2*) are sometimes called 'soft' interventions, and there is a burgeoning literature on how they can be used to learn about causal structures under certain assumptions (see Eberhardt and Scheines (2006), for example). But if what we want is an account of the *truth conditions* of

[14] C.f. Chapter Two, Section 3, above.

causal claims, something like (*I2*) is needed, because without it, a change in *Y* brought about by an intervention on *X* with respect to *Y* cannot be a necessary condition of *X*'s being a cause of *Y*, since a causal connection may exist between the two variables even though no change in *Y* occurs following the intervention *because no change in X occurs either*. Increasing parental income, for example, is not an acceptable intervention on whether or not a child attends private school, because the parents might have reasons for not sending their children to private school, such as political objections to private education, that have nothing to do with lack of funds. Here is the causal structure of this case again:

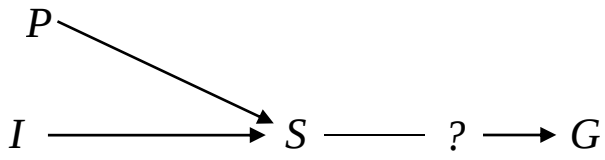

Fig. 5

Although *I* is a cause of *S* in this model, a change in *I* from 0 to 1 may fail to result in a change in the actual value of *S*, due to interference from *P*.

(*I2*) addresses this issue by stipulating that for *I* to be an intervention variable for *S*, *S* must cease to depend on the values of *every* variable except *I* when *I*=1. But this, I think, is also an overreaction. Consider the example Campbell discusses of an ordinary manipulation of an intention. Let *N*=1 if Jane intends to dance and 0 otherwise, *D*=1 if Jane dances and 0 otherwise, *B*=1 if Jane believes that dancing will make her happy and 0 otherwise, and *I*=1 if I tell Jane that dancing will make her happy and 0 otherwise. The causal model containing these four variables looks like this:

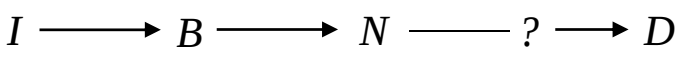

Fig. 6

Campbell correctly points out that *I* is not an intervention variable for *N* according to (*I2*), because when *I*=1, the value of *N* continues to depend on the value of a variable besides *I*, namely *B*. But notice the difference between fig.5 and fig.6 – in fig.5, *P* is part of a *separate* mechanism, the operation of which threatens to disrupt the mechanism connecting the intervention variable to the target variable. In fig.6, by contrast, *B* is *part* of the very mechanism through which the intervention operates. That the value of *N* continues to depend on the value of *B* in fig.6 is no surprise – it *has* to, otherwise *I*=1 wouldn't be an actual cause of the value of *N* in this model. Thus although I agree that *I* should not count as an intervention variable for *S* in fig.5, I see no reason why *I* should fail to count as an intervention variable for *N* in fig.6.

This suggests the following revision to the conditions on intervention variables – *I* is an intervention variable for *X* with respect to *Y* only if:

In every veridical model $\mathfrak{M}$ containing *I*, *X* and *Y*:

(*I1*) *I* is a variable-level cause of *X* in $\mathfrak{M}$;

(*I2**) When *I* takes its 'on' value, *I*=1 is an actual cause of the value of *X* along some path P, and no variable in $\mathfrak{M}$ not on P is an actual cause of the value of *X* in $\mathfrak{M}$;

(*I3*) Every path from *I* to *Y* in $\mathfrak{M}$ (if there is one at all) goes through *X*;

(*I4*) *I* is statistically independent of every variable in $\mathfrak{M}$ that is on a path to *Y* that does not go through *X*.

Both (*I2*) and (*I2**) imply that *I* is not an intervention variable for *S* in fig.5. But although (*I2*) implies that *I* is not an intervention variable for *N* in fig.6 either, it's perfectly consistent with (*I2**) that *I* is an intervention variable for *N* in fig.6, because even if the value of *B* is an actual cause of the value of *N* when *I* takes its 'on' value, *B* is on the path from *I* to *N* along which *I*=1 is an actual cause of the value of *N*. Let's call a manipulation of *X* satisfying (*I1*), (*I3*) and (*I4*) a *direct* intervention if it satisfies (*I2*),

and an *indirect* intervention if it satisfies (*I2**) but not (*I2*). My claim is that indirect interventions should be recognised as genuine interventions – the problem of abrupt transitions is a consequence of the fact that (*I2*) unnecessarily rules them out.

In defence of this claim, it's worth pointing out that very few (if any) manipulations actually carried out by practicing scientists – including manipulations of intentions – count as direct interventions. Consider, for instance, the studies reviewed by Webb and Sheeran's (2006) recent meta-analysis of the experimental evidence that intentions are causes of behaviour. Webb and Sheeran analysed 47 studies in which the effects of a particular 'intervention' on subjects' intentions to engage in a certain kind of behaviour – e.g. safe sex, smoking, visiting an internet site, and so on – and their subsequent behaviour is measured. The results show that the interventions had a sample-weighted average effect of size 0.66 on intentions and 0.36 on behaviour, showing, the authors conclude, that the correlation between intentions and behaviour is indeed causal. None of the 'interventions' in these studies satisfy (*I2*), however. For example, in Brubaker and Fowler (1990), college males were presented with "persuasive messages" on audiotape, in which a doctor challenges misconceptions about testicular self-examination (TSE), before urging the listener to carry out the procedure once a month. The intervention was found to have an effect on both intentions and behaviour. But Brubaker and Fowler explicitly acknowledge that the intervention affected the intentions of subjects *by* affecting the beliefs which constitute the rational causes of the forming of their intentions: "The experimental message…was designed to alter subjects' beliefs about the outcomes of performing TSE" (Brubaker and Fowler 1990: 1414). Hence the intervention failed to remove the target variable from under the influence of all its other causes, as (*I2*) requires. Needless to say, this apparent deficiency is not even remarked upon by the authors of the meta-analysis.

Of course, we shouldn't necessarily expect the interventionist account to analyse causation in terms of the kinds of manipulations that are *actually* performed in science

– all kinds of limitations usually make the ideal experiment impossible to perform, and so we should expect a certain amount idealization in an account of the truth-conditions of causal claims in terms of the experimental procedures used to test them.[15] But the real problem with (*I2*) is that it represents an ideal that scientists don't seem to feel any pressure to even *try* and approach. Typically, an experiment in the special sciences will measure the effect of a change in the intervention variable on both the target variable and the purported effect variable. Experimentalists will be concerned to ensure that, as best as possible, the value of the target variable can be uniquely determined by the value of the intervention variable. They *won't* typically be concerned to ensure that the intervention variable doesn't act on the target variable 'through' any other variable, however – *except*, of course, if there is reason to think that that intermediate variable is a confounder.

This latter concern sometimes seems to be what Campbell and Woodward have in mind when they motivate (*I2*):

> [S]uppose we leave intact the belief that φ-ing will make one happy. Then it is possible that the belief that φ-ing will make one happy causes both formation of the intention to φ and also directly causes performance of the action itself. In that case the intention to φ will be correlated with φ-ing even though the intention plays no role in causing the action. (Campbell 2007: 61-2; with minor changes in notation).

Similarly, Woodward insists on eliminating 'transition periods' between interventions and changes in the target variable, "because they may introduce factors that affect the effect independently of the putative cause" (Woodward 2003: 144-5).

But this possibility is *already* taken care of by (*I3*). For example, if there is a path from whether Jane believes that dancing will make her happy (*B*) to whether she dances (*D*)

[15] Again, the analogy with frequentism is helpful here – any plausible frequentist account should analyse objective probability in terms of *long-run* frequencies; longer, indeed, than anyone is in a practical position to determine.

that doesn't go through whether she intends to dance (*N*) in some model containing these three variables, then (*I3*) already implies that an intervention variable for *N* with respect to *D* shouldn't also be a cause of *B* in those models. If there is no such path, however, then there is no reason why an intervention on *N* with respect to *D* shouldn't also be a cause of *B*. In other words, there seems to be no reason why an intervention cannot act 'through' other variables, so long as those variables are not independent causes of the purported effect variable.

Another purported motivation for (*I2*), which Woodward mentions in passing, has to do with backward causation. According to Lewis, the closest possible world in which Tushar isn't in the outside lane at *t* is one in which a small 'miracle' occurs in his brain a few seconds before *t*, after which he smoothly transitions into the inside lane. But this seems to imply that, if Tushar hadn't been in the outside lane at *t*, a miracle would have occurred a few seconds before *t* – and it follows from *this*, Woodward argues, that on Lewis's counterfactual account of causation, "we get backward causation in a case in which backward causation is clearly not at work" (Woodward 2003: 143).[16] Woodward seems to think that (*I2*), by eliminating 'transition periods', also deals with this problem of widespread backwards causation.

But again, this issue is *already* taken care of by (*I3*). Consider fig.6 again:

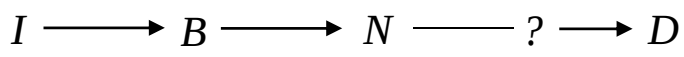

Fig. 6

[16] Lewis has a response to this objection, however, which Woodward doesn't acknowledge: "There may be a variety of ways the transition might go, hence there may be no true counterfactuals that say in any detail how the immediate past would be if the present were different. I hope not, since if there were a definite and detailed dependence, it would be hard for me to say why some of this dependence should not be interpreted – wrongly, of course – as backward causation over short intervals of time in cases that are not at all extraordinary." (Lewis 1979: 463).

It's consistent with (*I2**) that *I* is an intervention variable for *N* with respect to *D*. But it's not consistent with (*I3*) that *I* is an intervention variable for *N with respect to B*, because there is a path in this model from *I* to *B* which does not go through *N*. Thus it doesn't follow from the fact that a change in *I* from 0 to 1 results in a change in the value of *B* that *N* is a cause of *B* – in other words, it doesn't follow from the existence of a 'transition period' between an intervention and a change in the target variable that the change in the target variable is a cause of the earlier events in the transition period, contrary to what Woodward seems to think.

In summary, I think there are no good reasons to rule out indirect interventions. They are consistent with the original motivations for introducing the surgical constraint, they are frequently performed in science, and they don't lead to widespread backward causation. The problem of abrupt transitions is a consequence of the fact that (*I2*) is stronger than it needs to be.

7. Rational Causation Revisited

So far I have argued that (*I2**), and not (*I2*), is the correct constraint on interventions to adopt. This means that a manipulation of a subject's intentions which proceeds by way of telling her what will or won't make her happy may well count as an intervention, even though her intentions will continue to depend on their usual rational causes. And *this* means that our interest in the causal consequences of intentions is not necessarily an interest in the behaviour of agents lacking rational autonomy.

There are exceptions to this general rule, however, even after we replace (*I2*) with (*I2**). Suppose that Johnny is a secret agent, who forms a belief that there is poison in his wine. This causes him to form an intention to drop (and therefore smash) his glass, thereby avoiding having to drink the wine without arousing suspicion. However, whenever Johnny believes he is in danger he gets nervous, and this causes his palms to sweat which makes him more likely to drop whatever he's holding (he never did

successfully complete secret agent training). In this case, there is a path from whether or not Johnny believes that his wine is poisoned to whether or not he drops the glass, which doesn't go through whether or not Johnny intends to drop his glass. An intervention on Johnny's intention to drop his glass must therefore, given (*I3*), manipulate his intention without affecting whether or not he believes the wine is poisoned; that is, without affecting the rational causes of Johnny's intention. So in *some* cases, interventions involving the suspension of an agent's rational autonomy may be required to establish a causal connection between intentions and behaviour, albeit because of (*I3*) and not because of (*I2**). But these are special cases. Indeed, one could argue that cases like these strike us as odd or problematic precisely *because* of the weirdness of the interventions required to establish the causal claims in question. The point is that there is an important distinction between these kinds of cases and everyday cases of psychological causation, one over which the surgical constraint's blanket ban on indirect interventions runs roughshod, but which can be captured by my revised set of conditions on interventions.

8. Conclusion

In this paper, I've argued that the problem Campbell raises for interventionist approaches to causation in psychology is an instance of a wider problem, the problem of abrupt transitions, which arises from the so-called 'surgical constraint' on interventions. The correct solution to this problem is not to abandon the surgical entirely, as Campbell recommends, but rather to replace it with a weaker constraint, one which is consistent with the possibility of indirect interventions, but which nevertheless does all the work the surgical constraint was designed to do. On this revised version of interventionism, it doesn't follow – except perhaps in certain interesting special cases – that our interest in psychological causation is an interest in the behaviour of agents lacking rational autonomy.

References

Brubaker, R. G. and Fowler, C. (1990). Encouraging college males to perform testicular self-examination: Evaluation of a persuasive message based on the revised theory of reasoned action. *Journal of Applied Social Psychology 20*(17), 1411-1422.

Campbell, J. (2007). An interventionist approach to causation in psychology. In A. Gopnik and L. Schulz (eds.), *Causal learning: Psychology, philosophy, and computation*, Oxford: Oxford University Press.

Campbell, J. (2010). Independence of variables in mental causation. *Philosophical Issues 20*(1), 64-79.

Coleman, J. and Hoffer, H. (1987). *Public and private high schools*. New York: Basic Books.

Collingwood, R. (1940). *An essay on metaphysics*. Oxford: Clarendon Press.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Oxford University Press.

Edgington, D. (2011). Causation first: Why causation is prior to counterfactuals. In C. Hoerl, T. McCormack and S. R. Beck (eds.), *Understanding counterfactuals, understanding causation*, Oxford: Oxford University Press.

Franklin-Hall, L. (forthcoming). High-level explanation and the interventionist's variables problem. *British Journal for the Philosophy of Science.*

Gasking, D. (1955). Causation and recipes. *Mind 64*(256), 479-487.

Hausman, D. (1986) Causation and experimentation. *American Philosophical Quarterly 23*(2), 143-154.

Hesslow, G. (1976). Two notes on the probabilistic approach to causality. *Philosophy of Science 43*(2), 290-292.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy 98*(6), 273-299.

Hitchcock, C. (2012). Probabilistic causation. In E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy (Winter 2012 Edition)*, URL = <http://plato.stanford.edu/archives/win2012/entries/causation-probabilistic/>.

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Oxford: Oxford University Press.

Leslie, S. (2012). Generics. In G. Russell and D. Fara (eds.), *Routledge handbook of philosophy of language*, London: Routledge.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy 67*(13), 427-446.

Lewis, D. (1973). *Counterfactuals*. Oxford: Oxford University Press.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs 13*(4), 455-476.

Lewis, D. (1986). Causation. In D. Lewis, *Philosophical papers (vol. 2)*, Oxford: Oxford University Press.

Menzies, P. (2008). The exclusion problem, the determination relation, and contrastive causation. In J. Hohwy and J. Kallestrup (eds.), *Being reduced: New essays on reduction, explanation, and causation*, Oxford: Oxford University Press.

Menzies, P. and Price, H. (1993). Causation as a secondary quality. *British Journal for the Philosophy of Science 44*(2), 187-203.

Northcott, R. (2008). Causation and contrast classes. *Philosophical Studies 139*(1), 111-123.

O'Leary-Hawthorne, J. and Price, H. (1996). How to stand up for non-cognitivists. *Australasian Journal of Philosophy 74*(2), 275-292.

Pearl, J. (2000). *Causality: Models, reasoning and inference.* New York: Cambridge University Press.

Schaffer, J. (2016). Grounding in the image of causation. *Philosophical Studies 173*(1), 49-100.

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, prediction, and search (2nd ed.)*. Cambridge, MA: The MIT Press.

Wallace, D. (2012). *The emergent multiverse: Quantum theory according to the Everett interpretation*. Oxford: Oxford University Press.

Webb, T. L. and Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin 132*(2), 249-268.

Weslake, B. (forthcoming). Exclusion excluded. *International Studies in the Philosophy of Science.*

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

Woodward, J. (2008). Response to Strevens. *Philosophy and Phenomenological Research 77*(1), 193-212.

von Wright, G. (1971). *Explanation and understanding*. Ithaca, New York: Cornell University Press.

Chapter Five

Causal Contribution

Abstract: The practice of comparing 'degrees of contribution' of different causes is pervasive in many disciplines. Yet it has received little philosophical attention. In this paper, I argue that, although causation is not a scalar relation, it is a relation to which different events can contribute to different degrees. I then motivate a probabilistic analysis of an event's degree of contribution to a causing of an effect, showing how it has a number of advantages over previous accounts.

1. Introduction

It's often natural to compare two events by describing one as *more of a cause* of an effect than the other. A teacher might describe a student's lack of preparation as more of a cause of his poor exam performance than the difficulty of the questions, for example (Hart and Honoré 1985: 233). A historian might describe imperialism as more of a cause of the First World War than the rigid military mobilisation schedules. Or a policeman might describe the driver's drunkenness as more of a cause of a car crash than the adverse weather conditions. Similar talk of 'degrees of contribution', of 'causal potency' or 'causal efficacy', and of 'chief', 'main' or 'principal' causes, is pervasive in many disciplines, including the natural and social sciences, history and the law. Yet these kinds of comparisons have received scant attention in the philosophy literature. My aim in this paper is therefore to develop a metaphysics of causal contribution. I will argue that, although causation is not a scalar relation, it is nevertheless a relation to which different events can contribute to different degrees. I'll then motivate a novel analysis of an event's degree of contribution to a causing of

an effect in a probabilistic framework, showing how it has a number of advantages over alternative accounts. Given the close conceptual connections between causation and responsibility, an account of causal contribution promises to have a number of interesting applications, especially to ethics and the law – I explore one of these applications further in Chapter Six.

2. 'More of a Cause'

The locution 'more of a cause' has a number of importantly distinct meanings, which should be clearly distinguished from the outset. The first way to be 'more of a cause' of an effect is to cause a 'larger part' of that effect. If I have a contract with company A worth £100 and a contract with company B worth £400, and both A and B breach the terms of their contracts leaving me £500 out of pocket, there's an obvious sense in which B's breach was 'more of a cause' of my total losses than A's breach. This is because my total losses are easily *divisible* into discrete chunks, with A's actions causing one and B's actions causing the other, and the chunk caused by B's actions is 'larger' than the chunk caused by A's actions.

Another way to be 'more of a cause' of an effect is to make a *larger difference* to that effect. *A* made more of a difference to *P* than did *B* just in case "had *B* not occurred, something would have occurred which more closely approximates *P* than had *A* not occurred" (Martin 1989: 78). For example, suppose that, but for the rigid military mobilisation schedules in place in early 20th Century Europe, the war that occurred would have been shorter and less severe than the First World War actually was; but also that, but for the imperialist ambitions of European nations, there wouldn't have been a war in Europe in the early 20th Century at all. Then there's a sense in which imperialism was 'more of a cause' of the First World War than the rigid mobilisation schedules. This isn't because imperialism caused a 'larger part' of the First World War

than did the mobilisation schedules; it's because imperialism made more of a difference.[1]

But there is at least one other sense in which one event may be 'more of a cause' of an effect than another. To illustrate, consider the following case:

> *Committee*
>
> D_1, D_2, D_3 and D_4 are the members of an executive committee of a manufacturing company. The committee votes on motions, which are carried by a simple majority. Every committee member has one vote each, except D_1, the chair of the committee, who has two votes. The committee unanimously votes not to replace the company's outdated equipment. The equipment later malfunctions, injuring an employee.

The employee's injury in this case isn't divisible into four parts, with each committee member's actions causing one part each. Moreover, the actions of every member of the committee made exactly the same difference *vis-à-vis* the employee's injury, namely none – even if D_1 had voted to replace the equipment, the motion would still have passed with a majority, and the injury would still have occurred, more or less exactly in the way it actually did; and the same goes for D_2, D_3 and D_4. Nevertheless, I submit, there's an intuitive sense in which D_1's action was more of a cause of the injury than D_2's action – it made a *larger contribution* to bringing the injury about, because D_1 had two votes and D_2 only had one. It's *this* sense of 'more of a cause' that is the target of my analysis in this paper.

Some authors have doubted whether there really is a coherent concept here to analyse. "[T]he idea [of] relative causal contributions to an injury that is indivisible is seemingly oxymoronic", according to Barker and Steele (2015: 67). Pearson (1979: 345) argues

[1] The notion of difference-making "admits of degree in a rough and multi-dimensional way" (Lewis 2004: 92), since there are many dimensions along which two things can differ. This is made explicit in Northcott's measure of 'causal strength', which is relativized to a choice of variables (Northcott 2012, 2013).

that since "[c]ausation…exists or it does not…one does not speak of 'degrees of causation'". The sentiment is echoed by Wright (1988: 1146), who agrees that "[c]ausation…is not a matter of degree", since "[s]ome condition either was or was not a cause (in the proper scientific sense)". Even the most recent Restatement of Torts is "quite explicit" in its opinion that "[t]here are no degrees of factual cause".[2] I think these authors are making an important metaphysical mistake. As I argue in section 3, although causation is not a scalar relation, it is nevertheless a relation to which different events can *contribute* to different extents. The remainder of the paper defends a novel analysis of an event's degree of contribution to a causing of an effect.

3. The Metaphysics of Contribution

Consider the following sentence:

(1) Alice and Bob surrounded the tree.

(1) is ambiguous. Read *distributively*, it follows from (1) that Alice surrounded the tree and Bob also surrounded the tree. But the more natural reading is the *collective* one, according to which Alice and Bob surrounded the tree *together* (by joining hands around the tree, for example). These are distinct states of affairs. On the distributive reading the tree is surrounded twice; on the collective reading it is only surrounded once, even though it is surrounded by two people.

Exactly the same is true of (2):

(2) The driver's drunkenness and the rainstorm caused the car crash.

Read distributively, it follows from (2) that the drunkenness caused the crash and the rainstorm also caused the crash. On the more natural collective reading, however, the

[2] American Law Institute, *Restatement (Third) of Torts: Liability for Physical and Emotional Harm*, §26 cmt. j (2009).

drunkenness and the rainstorm caused the crash *together*. These are distinct states of affairs. On the distributive reading, the crash was caused twice over – it was *overdetermined*, to put it another way – whereas on the collective reading the crash was only caused once, even though it was caused by two events. Causation relates *pluralities* to individuals, in general; and a plurality of causes can collectively cause an effect without any one of the plurality individually causing it.

Some relations are *scalar*. Take *loving*, for example: I can love someone a lot, and I can love one person more than I love another. But surrounding is not a scalar relation. Consider the following sentences, for example:

(3) #Alice surrounded the tree a lot.

(4) #Alice surrounded the tree more than Bob did.

(3) and (4) sound odd, even ungrammatical.[3] How could Alice have surrounded the tree 'a lot'? Either she surrounded it or she didn't!

But now consider the following sentences:

(5) Alice contributed a lot to the surrounding of the tree.

(6) Alice contributed more than Bob to the surrounding of the tree.

(5) and (6) are perfectly grammatical; indeed, they would be true if Alice and Bob collectively surrounded the tree but Alice had longer arms, and therefore reached further around the tree, than Bob. So *surrounding* is all-or-nothing; but that's perfectly consistent with the possibility of different people *contributing* to a surrounding to different degrees.

[3] There is perhaps a reading of (3) according to which Alice has surrounded the tree *many times*; but this clearly doesn't show that surrounding is a scalar relation (compare: it doesn't follow from the natural reading of 'I love you a lot' that I love you many times).

Moore (2009: 275) has recently argued that "[c]ausation is a scalar relation". I think he is simply wrong about this. Consider the following sentences, for example:

(7) #The driver's drunkenness caused the crash a lot.

(8) #The driver's drunkenness caused the crash more than the rainstorm did.

(7) and (8) sound odd, even ungrammatical. How could the drunkenness have caused the crash 'a lot'? Either it caused the crash or it didn't!

But now consider the following sentences:

(9) The driver's drunkenness contributed a lot to the causing of the crash.

(10) The driver's drunkenness contributed more to the causing of the crash than the rainstorm did.

(9) and (10) are perfectly grammatical. So *causing* is all-or-nothing; but this is perfectly consistent with the possibility of different *causes* contributing to a causing of an effect to different degrees.[4]

Note that contributing *to a causing* of an effect is not the same as contributing *to the effect*. To (causally) contribute to an effect is to cause *a part* of that effect. B's breach of contract in the example above causally contributed to my total losses in virtue of causing a fraction of them. This isn't what's going on in the car crash case. To say that the driver's drunkenness and the rainstorm collectively caused the crash is *not* to say that the driver's drunkenness caused one part of the crash and the rainstorm caused a different part of the crash. The drunkenness didn't cause the crash, nor did it cause a

[4] Many causal verbs have the same structure. I can't *author* a book a lot, for example, but I can contribute a lot to the authoring of a book. One way I might do this is by *authoring a large part* of the book. But I might also have supplied the majority of the ideas or done the bulk of the research – in this case, I would have contributed a lot to the authoring of a book, even though there is no part of the book that I (individually) authored. I use *surrounding* as opposed to *authoring* as an analogy, primarily because it is plausibly part of the meaning of 'author' that a book can't be authored more than once, whereas it's generally agreed that an effect can be caused more than once (although see Unger 1977).

part of the crash; it contributed to a causing of the crash. Consider our surrounding analogy again: I can surround a tree by reaching all the way around it; I can surround *part* of a tree by reaching around one of its branches; and I can *contribute to a surrounding* of the tree by joining hands with another person around the tree. These are all distinct states of affairs; so it is with causation.

I claim to have isolated a concept – the concept of *contributing to a causing*. The next step is to get a bit more precise on what it is for an event to contribute to a causing of an effect to a particular degree. This is the aim of the next section.

4. Degrees of Causal Contribution

In Chapter One, I defended the following regimentation of the idea that causes *necessitate* their effects:

> (NC): $C_1,\ldots, C_n$ collectively caused E relative to **B** and **O** only if E occurs in every element of $\max(\mathbf{B}\cup\{C_1,\ldots, C_n\}, \mathbf{O})$.

According to (NC), causal claims are relativized to a modal base **B** and an ordering source **O** provided by the conversational context, which together pick out a restricted domain of possible worlds. There are contexts in which I can truly utter the sentence 'The short circuit caused the fire', for example – even though there are possible worlds in which the short circuit occurs without the fire – because in these contexts we are 'holding fixed' the fact that there was oxygen in the atmosphere, the fact that no macroscopic quantum-tunnelling events occurred, and so on.

I was keen to stress in Chapter One that (NC) is only a necessary condition on causation. One obvious reason why it wouldn't be a good idea to replace the 'only if' in (NC) with 'if and only if' is that, according to such a principle, every event would count as a cause of every other, at least relative to an empty ordering source; for if X is an

arbitrary event, and *E* occurs in every element of max(**B**∪{C_1,..., C_n}, ∅), then *a fortiori* *E* occurs in every element of max(**B**∪{C_1,..., C_n, X}, ∅).[5]

In light of this, some authors have advocated a stronger condition on causation – causes *non-redundantly* necessitate their effects (see Mackie 1965; Wright 1985; Strevens 2007). The idea is that if C_1,..., C_n collectively caused *E*, then C_1,..., C_n jointly necessitated *E and the same is not true of any proper sub-plurality of C_1,..., C_n*. More precisely:

> NON-REDUNDANT NECESSARY CONNECTIONS (NR-NC): C_1,..., C_n collectively caused *E* relative to **B** and **O** only if:
>
> – *E* occurs in every element of max(**B**∪{C_1,..., C_n}, **O**), and
>
> – for all proper sub-pluralities C_i,..., C_j of C_1,..., C_n, it's not the case that *E* occurs in every element of max(**B**∪{C_i,..., C_j}, **O**).

It will be useful for what follows to restate (NR-NC) in probabilistic terms. Let P(*p*) be the measure of *objective chance* over the set of propositions. By 'objective chance', I just mean what we normally take ourselves to mean when we say that the chance of a fair coin landing heads is 0.5. I don't take this to be a fact about any actual agent's credence function (although see Lewis 1980). Nor do I take it to imply indeterminism – the chance of a fair coin landing heads is 0.5, whether or not the laws of nature are deterministic. Probabilities of this kind are pervasive in science, most notably statistical mechanics. How exactly they should be understood is controversial,[6] but I won't take a stance on that question here.

Now consider the following principle:

---

[5] This conclusion doesn't follow relative to non-empty ordering sources, because although ∩(**B**∪{C_1,..., C_n, X}) is a proper subset of ∩(**B**∪{C_1,..., C_n}), max(**B**∪{C_1,..., C_n, X}, **O**) needn't be a proper subset of max(**B**∪{C_1,..., C_n}, **O**).

[6] See Wallace (2014) for an opinionated introduction.

(NR-NC*): $C_1,\ldots, C_n$ collectively caused E relative to **B** and **O** only if:

– $P(E \mid \max(\mathbf{B} \cup \{C_1,\ldots, C_n\}, \mathbf{O})) = 1$, and

– for all proper sub-pluralities $C_i,\ldots, C_j$ of $C_1,\ldots, C_n$,

$P(E \mid \max(\mathbf{B} \cup \{C_i,\ldots, C_j\}, \mathbf{O})) < 1$.

(NR-NC*) isn't *quite* a straightforward restatement of (NR-NC) – if there are an infinite number of propositions, for example, it could be that $P(E \mid \max(\mathbf{B} \cup \{C_1,\ldots, C_n\}, \mathbf{O})) = 1$ even if there is an element of $\max(\mathbf{B} \cup \{C_1,\ldots, C_n\}, \mathbf{O})$ in which E doesn't occur, so long as the set of such worlds is assigned measure zero by P. (By analogy: The probability of me losing a lottery with an infinite number of tickets is 1, even though there is some possible world in which I win.) But I'll ignore this complication here, since it won't be relevant in what follows.

If Alice and Bob collectively surround a tree, but Alice has longer arms, Alice contributed more than Bob to the surrounding of the tree. Intuitively this is because, although neither Alice nor Bob individually surrounded the tree, Alice in some sense came *closer* to surrounding the tree by herself than Bob did. I think a similar thing is true about causation. Suppose the driver's drunkenness and the rainstorm collectively caused the car crash, relative to some modal base **B** and ordering source **O**. Let D, R and E be the events of the driver's drunkenness, the rainstorm and the crash, respectively, occurring. Then according to (NR-NC), $P(E \mid \max(\mathbf{B} \cup \{D, R\}, \mathbf{O})) = 1$, whereas $P(E \mid \max(\mathbf{B} \cup \{D\}, \mathbf{O})) < 1$ and $P(E \mid \max(\mathbf{B} \cup \{R\}, \mathbf{O})) < 1$. But if $P(E \mid \max(\mathbf{B} \cup \{D\}, \mathbf{O})) > P(E \mid \max(\mathbf{B} \cup \{R\}, \mathbf{O}))$ – that is, if the probability of the crash occurring is greater conditional on the drunkenness occurring than it is conditional on the rainstorm occurring – then although neither the drunkenness nor the rainstorm individually caused the crash, the drunkenness in some sense came *closer* to causing the crash by itself than the rainstorm did, because it came closer to individually necessitating it. This, I think, is a situation in which it would be appropriate to say that the drunkenness contributed more to the causing of the crash.

Here's a natural way of cashing out this thought. Let $f(C_i, [C_1,\ldots,C_n]\rightarrow E)^{\mathbf{B},\mathbf{O}}$ be the function that returns C_i's degree of contribution to the causing of E by the plurality of events $C_1,\ldots, C_n$, relative to **B** and **O**. Then:

CAUSAL CONTRIBUTION (CC): If $C_1,\ldots, C_n$ collectively caused E relative to **B** and **O**, then $f(C_i, [C_1,\ldots,C_n]\rightarrow E)^{\mathbf{B},\mathbf{O}} = \dfrac{P(E \mid \max(\mathbf{B}\cup\{C_i\}, \mathbf{O}))}{\sum_{j=1}^{n} P(E \mid \max(\mathbf{B}\cup\{C_j\}, \mathbf{O}))}$

In words: The degree of contribution of a cause to a causing of an effect is equal to the probability (in the context) of the effect occurring conditional on the cause occurring, divided by the sum of the conditional probabilities for all the events involved in that causing. The denominator of this fraction is a renormalizing factor that ensures that degrees of contribution to a particular causing of E always sum to 1 (i.e. it follows from (CC) that $\sum_{i=1}^{n} f(C_i, [C_1,\ldots,C_n]\rightarrow E)^{\mathbf{B},\mathbf{O}} = 1$). This ensures that $f(C_i, [C_1,\ldots,C_n]\rightarrow E)^{\mathbf{B},\mathbf{O}}$ is not sensitive to the unconditional probability of E. In effect, (CC) simply *compares* how close each cause of E came to individually causing E, so that the uniform effect of P(E) on each conditional probability is cancelled out.

Three features of (CC) are worth emphasising. Firstly, the probabilities here are *not* epistemic probabilities; in particular, $P(E \mid \max(\mathbf{B}\cup\{C\}, \mathbf{O}))$ is *not* the epistemic probability that C caused E on our current evidence. If C *in fact* (individually) caused E relative to **B** and **O**, then $P(E \mid \max(\mathbf{B}\cup\{C\}, \mathbf{O})) = 1$, and so $f(C, [C]\rightarrow E) = \frac{1}{1} = 1$, regardless of how likely it is on our current evidence that C caused E.

Relatedly, (CC) isn't committed to any analysis of what it is to be *a cause* of an effect. In particular, it doesn't imply that C is a cause of E if (or only if) C raises the probability of E (i.e. if $P(E \mid \max(\mathbf{B}\cup\{C\}, \mathbf{O})) > P(E \mid \max(\mathbf{B}\cup\{\neg C\}, \mathbf{O}))$). (CC) simply provides a way of determining, *given* that C is a cause of E, C's degree of contribution to the causings of E to which it contributes. If C isn't a cause of E, then C didn't contribute to

any causing of E, and so its degree of contribution to every causing of E is undefined by (CC), regardless of the value of $P(E \mid \max(\mathbf{B} \cup \{C\}, \mathbf{O}))$.

Finally, you'll notice that degrees of causal contribution are relativized to a modal base and an ordering source, according to (CC). Evaluations of an event's degree of contribution to a causing of an effect are always made relative to a set of facts we're 'holding fixed' and a way of 'ranking' possible worlds, in the way described in Chapter One. Generally speaking, the less we include in the modal base and ordering source, the less an event will contribute to causings of an effect. This is because the more accessible worlds there are, the larger a plurality of events needs to be in order to count as necessitating an effect, and so the more causes there will be among which to distribute degrees of contribution. The driver's drunkenness might contribute to degree 0.6, say, to the causing of the car crash relative to the context in which the crash was collectively caused by the drunkenness and the rainstorm, but it will contribute *much* less to the causing of the crash relative to the context in which the crash was collectively caused by the drunkenness, the rainstorm, the design of the tyres, the camber of the road, the absence of macroscopic quantum-tunnelling events, and so on. For much of the remainder of this paper, I will drop reference to **B** and **O** when it is clear which context is at issue; but it's important to remember that whenever we evaluate an event's degree of contribution to a causing of an effect, we are always (even if implicitly) doing so relative to a particular set of accessible worlds.

Now let's see how (CC) works by applying it to our *Committee* example above. Recall that the committee consists of four members, D_1, D_2, D_3 and D_4, where D_1 has two votes and the others have one vote each. They all vote in favour of a motion which results in an injury to an employee. Let V_1, V_2, V_3 and V_4 be the events of D_1, D_2, D_3 and D_4, respectively, voting in favour, and let H be the event of the injury occurring. There are a total of five votes, so the minimum number of votes needed for a majority is three. None of V_1, V_2, V_3 or V_4 individually necessitated H. But relative to a suitable modal

base (one which implies that the injury will occur if, but only if, the motion passes), four pluralities of events jointly and *non-redundantly* necessitated H: V_1 and V_2, V_1 and V_3, V_1 and V_4, and V_2, V_3 and V_4 (each of these pluralities correspond to three votes being cast in favour). So according to (NR-NC), the injury was caused *four times* by four different (though overlapping) pluralities of events. This is the sense in which the injury was overdetermined.

We can now calculate the degrees of contribution each vote made to each of the four causings of the injury. To keep things simple, let's assume that each committee member has a probability (in the relevant context) of 0.5 of voting in favour of the motion, regardless of the actions of the others (we'll revisit this assumption below). Then $P(H \mid V_1)$ – the probability of the injury occurring conditional on D_1 voting in favour – is equal to the probability of one or more of the remaining three committee members voting in favour. This is given by the sum of binomial coefficients $\sum_{i=1}^{3}({}^3C_i) = 0.875$. $P(H \mid V_2)$, by contrast, is equal to the probability of D_1 voting in favour, plus the probability of D_3 and D_4 both voting in favour, minus the probability of all three voting in favour (to avoid double-counting), which is $0.5 + 0.5^2 - 0.5^3 = 0.625$. By similar reasoning, $P(H \mid V_3) = P(H \mid V_4) = 0.625$. Plugging these values into (CC), we get the following results:

$$f(V_1, [V_1,V_2]\rightarrow H) = f(V_1, [V_1,V_3]\rightarrow H) = f(V_1, [V_1,V_4]\rightarrow H) = \frac{0.875}{0.625 + 0.875} \approx 0.58$$

$$f(V_2, [V_1,V_2]\rightarrow H) = f(V_3, [V_1,V_3]\rightarrow H) = f(V_4, [V_1,V_4]\rightarrow H) = \frac{0.625}{0.625 + 0.875} \approx 0.42$$

$$f(V_2, [V_2,V_3,V_4]\rightarrow H) = f(V_3, [V_2,V_3,V_4]\rightarrow H) = f(V_4, [V_2,V_3,V_4]\rightarrow H) = \frac{0.625}{3 \times 0.625} = \frac{1}{3}$$

In other words, V_2, V_3 and V_4 made equal contributions to their causing of the injury – this is what we expected, since D_2, D_3 and D_4 have one vote each – but V_1 contributed *more* than V_2, V_3 and V_4 to the other causings of the injury – which is also what we expected, since D_1 has more votes than the others.

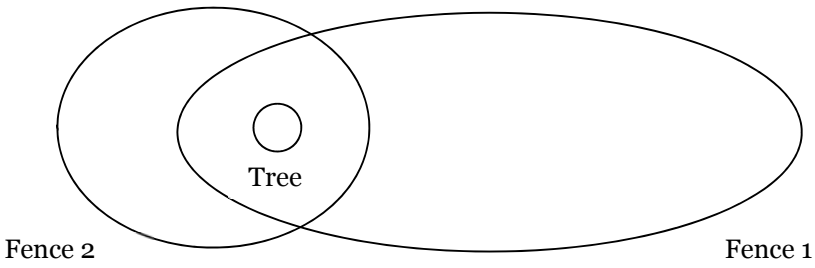

Fig. 1

Imagine a tree surrounded by two wooden fences, which intersect at two points (see fig. 1). Consider the plank of wood at one of the points of intersection. It contributes to two different surroundings of the tree, the surrounding by Fence 1 and the surrounding by Fence 2. Moreover, it contributes *less* to the surrounding by Fence 1 than it does to the surrounding by Fence 2, since Fence 1 is longer than Fence 2. A plank of wood can contribute to some degree to *one* surrounding of a tree and contribute to a *different* degree to another surrounding of that very same tree. The same is true for causation – an event can contribute to some degree to *one* causing of an effect and contribute to a *different* degree to another causing of that very same effect. In *Committee*, for example, V_2 contributes to one causing of the injury to degree 0.42 and contributes to a different causing of the injury to degree $\frac{1}{3}$.

I've been assuming so far that there are no correlations between voters. Suppose instead that D_2 is particularly *influential*, so that the others are more likely to vote in favour if she does the same. Suppose, to be more specific, that the probability (in the relevant context) of D_i voting in favour conditional on D_2 voting in favour is 0.75, for all $i \neq 2$, whereas the probability of D_2 voting in favour conditional on D_i voting in favour is still just 0.5, for all $i \neq 2$. Then V_2's degrees of contribution to the causings of the injury are given as follows:

$$f(V_2, [V_1,V_2]\rightarrow H) = \frac{0.75 + 0.75^2 - 0.75^3}{\sum_{i=1}^{3}({}^3C_i) + (0.75 + 0.75^2 - 0.75^3)} \approx 0.51$$

$$f(V_2, [V_2,V_3,V_4]\rightarrow H) = \frac{0.75 + 0.75^2 - 0.75^3}{(0.75 + 0.75^2 - 0.75^3) + 2(0.5 + 0.5^2 - 0.5^3)} \approx 0.42$$

Hence if we make D_2 more influential, her degrees of contribution to the causings of the injury increase, according to (CC) – from 0.42 to 0.51, and from ⅓ to 0.42.

What happens if we add more members to our committee? Suppose for example that we add D_5 and D_6 to make a six-member committee. Let's also assume that D_5 and D_6 have one vote each, so that D_1 is still the only committee member with two votes. Now there are seven votes available, so four votes are needed for a majority. This means that there are *fourteen* causings of the injury, one for each plurality of events corresponding to four votes being cast in favour. V_1 contributes to ten of these causings (the causing by V_1, V_2 and V_3 is one example). What is V_1's degree of contribution to these causings of the injury? Assuming again that the probability (in the relevant context) of any one committee member voting in favour is 0.5, regardless of the actions of the others, $P(H \mid V_1)$ is equal to the probability of two or more of the remaining five committee members voting in favour, which is $\sum_{i=2}^{5}({}^5C_i) = 0.8125$. By similar reasoning, $P(H \mid V_2) = P(H \mid V_3) = \sum_{i=3}^{4}({}^4C_i) + 6(0.5^3) - 4(0.5^4) - 0.5^5 = 0.78125$. So the degrees of contribution to the causing of the injury by V_1, V_2 and V_3 are given as follows:

$$f(V_1, [V_1,V_2,V_3]\rightarrow H) = \frac{0.8125}{0.8125 + (2 \times 0.78125)} \approx 0.34$$

$$f(V_2, [V_1,V_2,V_3]\rightarrow H) = f(V_3, [V_1,V_2,V_3]\rightarrow H) = \frac{0.78125}{0.8125+(2\times 0.78125)} \approx 0.33$$

If we increase the number of committee members by two, then, the number of causings to which V_1 contributes goes up (from three to ten), but the degree of contribution it makes to those causings goes down (from about 0.58 in the original case to about 0.34

in this case). In general, as the number of committee members tends to infinity, the number of causings of the injury to which each vote contributes tends to infinity, and its degree of contribution to those causings tends to zero. Notice also that the difference between V_1's and V_2's degrees of contribution is much smaller in this case than in the original case (about 0.01 compared to 0.16) – D_1's extra vote becomes less and less important as the number of committee members, and therefore the number of available votes, increases.

5. Comparisons and Alternatives

Although the concept of causal contribution has received comparatively little attention in the philosophical literature thus far, a few measures of 'causal strength' or 'degrees of causation' have been proposed before. In this section I'll compare (CC) to these alternatives, arguing that none of them succeed in capturing the concept (CC) purports to analyse.

Some authors, even when it is clear they have something more subtle in mind, talk of the 'causal contribution' of an event in terms of the "the *part* or *proportion* of an outcome that [the event] causes" (Bernstein forthcoming).[7] As explained above, (CC) is a measure of an event's contribution *to a causing* of an effect, which is not the same as the 'part' or 'proportion' of the effect caused by the event – V_1 didn't cause the injury in *Committee*, nor did it cause a 'part' of the injury, but it did contribute to (several) causings of the injury, and (CC) can determine V_1's degrees of contribution to these causings.

Other authors talk of 'causal contribution' in terms of the *difference* a cause made with respect to an effect. According to Sober (1988: 303), for example, "the contribution a

[7] Tadros (ms) also identifies "a sense in which there are degrees of causal contribution to a threat", according to which "X *v*ing can causally contribute to the threat without causing the whole threat…[i]t can do so by causing part of the threat".

cause makes and the difference it makes seem to be one and the same issue". In a series of papers, Northcott (2005, 2006, 2008a, 2008b, 2012, 2013) defends a measure of 'relative causal strength' (sometimes 'relative causal efficacy' or 'degrees of causation'), where "the strength of a cause is how much difference it makes" (Northcott 2013: 3090).[8] The concept of difference-making is certainly an important one, and it certainly comes in degrees. But it is not the target of this paper. None of the votes in *Committee* made any difference whatsoever to whether, or to how, the injury occurred; and yet there is still intuitively a sense in which D_1's vote contributed more to bringing the injury about than D_2's vote did. (CC) is an attempt to capture *this* concept of causal contribution.

Fitelson and Hitchcock (2011) put together a helpful survey of a number of 'probabilistic measures of causal strength' from the philosophical and scientific literatures, some of which bear some resemblance to (CC). Consider for example what they call the 'Suppes measure' of causal strength, inspired by Suppes's (1970) probabilistic account of causation.[9] Let $s(C, E)$ be the function that returns C's degree of 'causal strength' with respect to E on the Suppes measure. Then:

THE SUPPES MEASURE (SUP): If C is a cause of E, $s(C, E) = P(E \mid C)$

One problem with (SUP) (as Fitelson and Hitchcock recognise) is that $s(C, E)$ is sensitive to the unconditional probability of E. If E is very likely to occur *regardless* of whether C occurs, $s(C, E)$ will be close to 1, even if C intuitively played a relatively minor role in bringing E about. As explained above, (CC) does not have this

[8] Lewis's (2004) account of 'causation as influence' also lends itself to a definition of causal difference-making, since "[i]nfluence admits of degree" (Lewis 2004: 92). An interesting variant on the difference-making idea is Chockler and Halpern's analysis of 'degrees of responsibility', where "the degree of responsibility of A for B" is a function of "the minimal number of changes that have to be made to obtain a contingency where B counterfactually depends on A" (Chockler and Halpern 2004: 94).

[9] Interestingly, Sprenger (ms) produces a representation theorem purporting to show that all measures of 'causal strength' intended to capture "to what extent the presence of C is *sufficient* for producing E" are ordinarily equivalent to the Suppes measure.

consequence, because (CC) only *compares* the conditional probabilities for each event involved in a causing, so that the uniform effect of P(E) on the conditional probabilities is cancelled out.

Relatedly, Rizzo and Arnold (1980) propose a theory of 'causal apportionment' in the context of the economic theory of tort law, which also bears some resemblance to (CC). Where C_1 and C_2 are causes of E, two quantities α and β are defined by Rizzo and Arnold as follows:[10]

$$\alpha = P(E \mid C_1)$$
$$\beta = P(E \mid C_2)$$

Then:

> THE RIZZO AND ARNOLD MEASURE (RA): "The *ratio*…(α/β) represents the relative contribution of one cause [C_1] with respect to the other [C_2]" (Rizzo and Arnold 1980: 1409).

Rizzo and Arnold go on to argue that damages in tort cases with two defendants should be apportioned according to the 'relative contribution' of one defendant's wrongdoing with respect to the other.

(RA) has a number of advantages over (SUP). Since (RA) only considers the *ratio* of conditional probabilities, it is insensitive to the unconditional probability of E. Indeed, the ratio α/β is equivalent to the ratio $f(C_1, [C_1,C_2]\rightarrow E) / f(C_2, [C_1,C_2]\rightarrow E)$ in the special case where C_1 and C_2 collectively caused E. But (CC) is also more discriminating than (RA). (RA) can only compare the relative contributions of *two* causes, whereas (CC) can straightforwardly handle cases with more than two causes. And if C_1 or C_2 contributes to more than one causing of E, (CC) can distinguish its degrees of

[10] I've omitted Rizzo and Arnold's reference to "standard environmental conditions" (Rizzo and Arnold 1980: 1409).

contribution to each of these causings, whereas (RA) cannot. As I argue below (Chapter Six, Section 5), these distinctions matter for Rizzo and Arnold's stated aim of determining the right way to apportion damages in tort cases with multiple defendants.

One final measure of 'degrees of causation' worth considering is the one recently proposed by Braham and van Hees (2009), which, like (CC), is inspired by Wright's (1985) account of causes as 'NESS conditions'. The measure is formulated within a game-theoretic framework and the details are somewhat complex, but the basic idea is easy to state. Let $\#(C, E)$ be the function that returns the *number* of causings of E to which C contributes. Then $bh(C, E)$, the 'degree of causal contribution of C to E' on Braham and van Hees's measure, is defined as follows:

THE BRAHAM AND VAN HEES MEASURE (BH): Where $C_1,\ldots, C_n$ are the causes of E,

$$bh(C_i, E) = \frac{\#(C_i, E)}{\sum_j \#(C_j, E)}$$

In *Committee*, for example, V_1 contributes to *three* causings of the employee's injury whereas V_2, V_3 and V_4 only contribute to two causings each. So $bh(V_1, H) = \frac{3}{9}$ and $bh(V_2, H) = bh(V_3, H) = bh(V_4, H) = \frac{2}{9}$. Interestingly, Braham and van Hees go on to claim that bh is equivalent to either the Public Good Index (Holler 1982) or the normalised Penrose-Banzhaf Index (Penrose 1946; Banzhaf 1965) of voting power, depending on which of two disambiguations of the NESS account one adopts.

Although (BH) can capture one important difference between D_1 and the other committee members in *Committee*, (CC) can do more than simply compare the *number* of causings to which an event contributes – it can also determine, for each one of these causings, that event's degree of contribution *to that causing*. This makes (CC) more powerful than (BH). To see this, note that (BH) only delivers interesting results

in cases of overdetermination. For example, if the car crash (E) was collectively caused by the driver's drunkenness (D) and the rainstorm (R), and E was not overdetermined, then D and R both contributed to a single causing of E, and so $bh(D, E) = bh(R, E) = 0.5$. Insofar as there is a sense in which the drunkenness might have been 'more of a cause' of the crash than the rainstorm, then, (BH) is incapable of capturing it. Indeed, Braham and van Hees themselves admit that, in such a case, (BH) designates both events as having made an equal contribution, "irrespective of their 'size'" (Braham and van Hees 2009: 330); which suggests that they have a concept of the 'size' of a cause in mind that their measure does not capture. I agree – (CC) is designed to capture this concept.

In summary, once we distinguish between different senses in which an event may be 'more of a cause' than another, and once we distinguish the idea of contributing *to a causing* of an effect from the distinct idea of (causally) contributing *to an effect*, it's clear that (CC) has a number of advantages over previous measures of 'degrees of causation' in the literature.

6. Spurious Correlations

There is one big problem with (CC) as it stands, however, which to some extent it shares with every probabilistic analysis of a causal concept – it goes awry in cases involving so-called 'spurious correlations'. Suppose that D_3 in *Committee* is happiest in winter, whereas the other committee members are happiest in summer. As a result, D_3 is more likely to vote in favour of any motion in winter, whereas the other committee members are more likely to do so in summer. Then, if we include these facts in the modal base, it follows that $P(E \mid \max(\mathbf{B} \cup \{V_3\}, \mathbf{O})) < P(E \mid \max(\mathbf{B} \cup \{V_2\}, \mathbf{O}))$ – the probability of the injury occurring will be less conditional on D_3 voting in favour than it will conditional on D_2 voting in favour, even though D_2 and D_3 have one vote each – because if D_2 is voting in favour it's probably summer and so the other

committee members will probably be voting in favour too, whereas if D_3 is voting in favour it's probably winter and so the other committee members probably won't be voting in favour. Given this, it follows from (CC) that $f(V_2, [V_2,V_3,V_4]\rightarrow H)^{\mathbf{B,O}} > f(V_3, [V_2,V_3,V_4]\rightarrow H)^{\mathbf{B,O}}$ – in other words, D_2's vote contributes more, in this context, to bringing the injury about than D_3's vote. This seems like the wrong result. Although the injury is more likely to occur if D_2 votes in favour than it is if D_3 votes in favour, we intuitively don't want this 'spurious' correlation to be reflected in the degrees of contribution made by V_2 and V_3 to their causings of the injury. Notice the difference between this case and the case where D_2 was particularly *influential* – in *that* case, it did seem right that V_2 should contribute more to the causings of the injury, because D_2's voting in favour was a *cause* of the other committee members voting in favour. This isn't what's going on in the present case; rather, what we have here is a case where D_2's voting in favour and the other committee members' voting in favour have a *common* cause, namely the fact that it's summer.

The right way to solve this problem, I think, is to adopt a proposal of Fenton-Glynn's (forthcoming) and replace conditional probabilities with *interventionist* conditional probabilities. Using some familiar notation introduced by Pearl (2000), let do(C, E) be the proposition, not that C occurs, but that C is *caused* to occur *by an intervention with respect to* E. The notion of an intervention was defined in more detail in Chapter Four (see Section 2), but intuitively, one can think of an intervention on C with respect to E as the event of "lifting [C] from the influence of the old functional mechanism and placing it under the influence of a new mechanism…while keeping all other mechanisms undisturbed" (Pearl 2000: 70). An intervention on the reading of a barometer with respect to the weather, for example, might involve opening up the barometer and fiddling about with the needle, thereby placing the barometer reading under the sole influence of a new mechanism without affecting the existing mechanism

connecting atmospheric air pressure to weather outcomes. The idea, then, is that the problem of spurious correlations can be solved by modifying (CC) as follows:

CAUSAL CONTRIBUTION* (CC*): If $C_1,\ldots, C_n$ collectively caused E relative to $\mathbf{B}$ and $\mathbf{O}$, then $f(C_i, [C_1,\ldots,C_n]\rightarrow E)^{\mathbf{B},\mathbf{O}} = \dfrac{P(E \mid \max(\mathbf{B}\cup\{do(C_i, E)\}, \mathbf{O}))}{\sum_{j=1}^{n} P(E \mid \max(\mathbf{B}\cup\{do(C_j, E)\}, \mathbf{O}))}$

For example, when calculating the degree of contribution of D_2's vote to the causings of the injury in *Committee*, we need to consider, not the probability of the injury occurring conditional on D_2 voting in favour of the motion, but rather the probability of the injury occurring conditional on D_2 being *caused* to vote in favour *by an intervention* with respect to the injury. Such an intervention might involve holding a gun to D_2's head and ordering him to vote in favour, for example. Assuming there are no correlations between such events and the voting patterns of the other committee members, the probability of the injury occurring conditional on D_2 being forced to vote in favour in this way will be no greater than the probability of the injury occurring conditional on D_3 being forced to vote in favour in the same way; so (CC*) correctly predicts that V_2 and V_3 contribute to the same degree to their causing of H, even despite the spurious correlation between the committee members' voting patterns. The intervention effectively 'breaks' the connection between D_2's voting in favour and the time of year, the common cause responsible for the spurious correlation. Of course, as soon as we appeal to the notion of an intervention, we're giving up on the prospect of a *reductive* analysis of degrees of causal contribution, a definition of an event's degree of contribution to a causing of an effect in entirely non-causal terms. But this arguably shouldn't come as a surprise – it's generally accepted that one cannot simply 'read off' causal facts from purely statistical information.[11]

[11] In practice statisticians and data scientists assume the Causal Markov Condition as a starting point (see Hausman and Woodward 1999, 2004), but there are circumstances in which this condition fails, and "it is a plausible conjecture that in these circumstances, no

7. Conclusion

I've argued in this paper that causation, though not a scalar relation, is a relation to which multiple events can contribute to different degrees. I've also motivated a probabilistic measure of an event's degree of contribution to a causing of an effect, and argued that it has a number of advantages over previous measures of 'degrees of causation' or 'causal strength' in the literature. Finally, I suggested a way of modifying the account using interventionist resources to deal with the possibility of spurious correlations.

The theoretical framework introduced in this paper has a number of important potential applications. 'Degrees of causation' have been invoked as part of proposed reforms to the law of torts with multiple defendants (see Chapter Six, Section 6, below), the mechanism of secondary liability in the criminal law (Moore 2009: 299-303), orthodox just war theory (Tadros ms), and other positions in ethics and legal theory. A rigorous metaphysics of causal contribution can help us better evaluate these proposals.

References

Banzhaf, J. F. (1965). Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review 19*(2), 317-344.

Barker, K. and Steele, J. (2015). Drifting towards proportionate liability: Ethics and pragmatics. *The Cambridge Law Journal 74*(1), 49-77.

Bernstein, S. (forthcoming). Causal proportions and moral responsibility. In D. Shoemaker (ed.), *Oxford studies in agency and responsibility (vol. 4)*, Oxford: Oxford University Press.

general test for causation in terms of conditional independence relationships will work" (Woodward 2003: 64).

Braham, M. and van Hees, M. (2009). Degrees of causation. *Erkenntnis 71*(3), 323-344.

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

Fenton-Glynn, L. (forthcoming). A proposed probabilistic extension of the Halpern and Pearl definition of 'actual cause'. *British Journal for the Philosophy of Science.*

Fitelson, B. and Hitchcock, C. (2011). Probabilistic measures of causal strength. In P. McKay Illari, F. Russo and J. Williamson (eds.), *Causality in the sciences*, Oxford: Oxford University Press.

Hart, H. L. A. and Honoré, T. (1985). *Causation in the law (2nd ed.)*. Oxford: Clarendon Press.

Hausman, D. M. and Woodward, J. (1999). Independence, invariance and the causal Markov condition. *British Journal for the Philosophy of Science 50*(4), 521-583.

Hausman, D. M. and Woodward, J. (2004). Modularity and the causal Markov condition: A restatement. *British Journal for the Philosophy of Science 55*(1),147-161.

Holler, M. J. (1982). Forming coalitions and measuring voting power. *Political Studies 30*(2), 262–271.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (ed.), *Studies in inductive logic and probability*, Berkeley: University of California Press.

Lewis, D. (2004). Causation as influence. In N. Hall, L. A. Paul and J. Collins (eds.), *Causation and counterfactuals*, Cambridge, MA: The MIT Press.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly 2*(4), 245-264.

Martin, R. (1989). *The past within us: An empirical approach to philosophy of history*. Princeton: Princeton University Press.

Moore, M. S. (2009). *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford: Oxford University Press.

Northcott, R. (2005). Pearson's wrong turning: Against statistical measures of causal efficacy. *Philosophy of Science 72*(5), 900-912.

Northcott, R. (2006). Causal efficacy and the analysis of variance. *Biology and Philosophy 21*(2), 253-276.

Northcott, R. (2008a). Can ANOVA measure causal strength?. *Quarterly Review of Biology 83*(1), 47-55.

Northoctt, R. (2008b). Weighted explanations in history. *Philosophy of the Social Sciences 38*(1), 76-96.

Northcott, R. (2012). Partial explanations in social science. In H. Kincaid (ed.), *Oxford handbook of philosophy of social science*, Oxford: Oxford University Press.

Northcott, R. (2013). Degree of explanation. *Synthese 190*(15), 3087-3105.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Pearson, R. N. (1979). Apportionment of losses under comparative fault laws – An analysis of the alternatives. *Louisiana Law Review 40*(2), 323-372.

Penrose, L. S. (1946). *Journal of the Royal Statistical Society 109*(1), 53-57.

Rizzo, M. J. and Arnold, F. S. (1980). Causal apportionment in the law of torts: An economic theory. *Columbia Law Review 80*(7), 1399-1429.

Sober, E. (1988). Apportioning causal responsibility. *Journal of Philosophy 85*(6), 303-318.

Strevens, M. (2007). Mackie remixed. In J. K. Campbell, M. O'Rourke and H. S. Silverstein (eds.), *Causation and explanation*, Cambridge, MA: The MIT Press.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland Publishing Company.

Tadros, V. (ms). Causal contributions and liability.

Unger, P. (1977). The uniqueness in causation. *American Philosophical Quarterly 14*(3), 177-188.

Wallace, D. (2014). Probability in physics: Statistical, stochastic, quantum. In A. Wilson (ed.), *Chance and temporal asymmetry*, Oxford: Oxford University Press.

Wright, R. W. (1985). Causation in tort law. *California Law Review 73*(6), 1735-1828.

Wright, R. W. (1988). Allocating liability among multiple responsible causes: A principled defense of joint and several liability for actual harm and risk exposure. *University of California at Davis Law Review 21*(4), 1141-1211.

Chapter Six

Partial Liability

Abstract: In most cases, liability in tort law is all-or-nothing. By contrast, I argue that a defendant should be held liable for a claimant's losses only to the degree to which the defendant's wrongdoing contributed to the causing of the claimant's harm. Using the analysis of causal contribution developed in Chapter Five, I defend novel interpretations of the legal concept of 'proximate causation' and the mechanisms for 'apportioning' damages in cases with multiple wrongdoers. I ground my approach in a conception of tort law as a system of corrective justice, which exists to correct injustices inflicted by one person on another, and *not* to redistribute the costs of harms or to punish wrongdoers.

1. Introduction

In most cases, liability in tort law is *all-or-nothing* – a defendant is either fully liable or not at all liable for the losses arising from a claimant's harm. By contrast, this paper defends a causal theory of *partial* liability. I will argue that a defendant should be held liable for a claimant's losses only to the degree to which the defendant's wrongdoing contributed to the causing of the claimant's harm.

I begin, in section 2, with a brief introduction to the corrective justice approach to tort law, according to which tort law is a mechanism for enforcing second-order corrective duties arising out of breaches of first-order duties of care. In section 3, I examine the law's fraught distinction between 'factual' and 'proximate' causation and criticise attempts to make sense of it in terms of the so-called 'relational account of wrongfulness'. Instead, as I argue in section 4, the concept of a 'proximate' cause can

and should be analysed in causal terms, as an event which contributed to a causing of an effect *to a significant degree*. In section 5, however, I argue that proximate causation doctrines ought to be replaced with a more fine-grained approach which recognises the possibility of *partial* liability for losses. I show how this approach can explain various statutory reforms to the law of cases involving multiple wrongdoers. Finally, in section 6, I consider some common objections to recognising partial liability, and reply to them by distinguishing the goals of corrective justice from the goals of *distributive* and *retributive* justice, and defending the controversial view that normative coherence in tort law can be achieved only if it is used to pursue purely corrective goals.

2. What is Tort Law?

According to Steel (2015: 10), "'tort law' refers primarily to the bodies of law which provide persons with remedies which respond to breaches of legal duties owed to them by other persons, where those duties are owed independently of contracts between them or declarations of trusts". This conception of tort law is often referred to as the *corrective justice* theory. It received its classic formulation in Book V of Aristotle's *Nicomachean Ethics* and forms the basis of many modern theories of tort (e.g. Coleman 1992; Ripstein 1998; Weinrib 1995). On the corrective justice theory, all legal persons bear certain first-order duties to other legal persons, who, correlatively, have certain rights against those duties being breached (Hohfeld 1917). If D breaches a first-order duty she bears to P, P has thereby been *wronged* by D. This triggers a second-order duty on the part of D, a duty to *correct* or *make right* that wrong. The point of tort law, according to the corrective justice theory, is to provide a mechanism for enforcing these second-order duties.

The standard characterization of the first- and second-order duties of tort law can be stated as follows (see Steel 2015: 39-42):

WRONGFUL CONTRIBUTION (WC): Where D and P are legal persons:

- D bears a first-order duty to P not to wrongfully contribute to a causing of harm to P, and
- if D breaches a first-order duty to P, D acquires a second-order duty to restore P to the state she would have been in, but for the harm.[1]

To illustrate, consider the following three cases:

Fire 1

Big Corp fails to replace its outdated equipment. The equipment short circuits, starting a fire that burns down Greta's house. Had the equipment been replaced, the short circuit wouldn't have occurred.

Fire 2

Big Corp fails to replace its outdated equipment. Luckily, the equipment doesn't short circuit. Unfortunately, lightning strikes a tree nearby, starting a fire that burns down Greta's house.

Fire 3

Big Corp buys new equipment with a high safety rating. Unfortunately, one of the new pieces of equipment short circuits unexpectedly, starting a fire which burns down Greta's house.

In *Fire 1*, Big Corp's activities caused the damage to Greta's house (relative to a natural modal base – see Chapter One). Moreover, those activities were *wrongful* – in failing to replace the outdated equipment, Big Corp failed to exercise the care that might reasonably be expected of it. Big Corp has therefore breached the first-order duty it bears to Greta not to wrongfully contribute to a causing of harm to her. This triggers a second-order duty on the part of Big Corp to restore Greta to the state she would have

[1] The word 'harm' here should be understood broadly to include physical changes in a person's body or property, reduction or non-receipt of financial assets, loss of freedom, mental suffering, or even loss of a valuable chance (see Steel 2015: 292-325) or interference with exclusive possession (c.f. the English tort of trespass).

been in, but for the damage to her house. This would normally be achieved through the paying of monetary damages to Greta to compensate her for the cost of repair, lost earnings, and so on. I'll say that Big Corp is *liable for Greta's losses*, in such a case.

By contrast, Big Corp is not liable for Greta's losses in either *Fire 2* or *Fire 3*, according to (WC). In *Fire 2*, although Big Corp acted negligently, its activities didn't contribute to a causing of the damage to Greta's house; and in *Fire 3*, although Big Corp's activities caused the damage to Greta's house, those activities weren't wrongful, since Big Corp took all reasonable precautions. So in neither case did Big Corp breach its duty to Greta not to wrongfully contribute to a causing of harm to her. Greta's losses in *Fire 2* and *Fire 3* must therefore be absorbed by Greta's insurance provider, or by the State, or else by Greta herself.

(WC) is controversial, both as a descriptive account of what tort law *is* and as a prescriptive account of tort law *ought* to be. Some proponents of the corrective justice approach to tort law have argued that, at least in some cases, defendants should be held liable for losses arising from all harms to which their actions causally contributed, *regardless* of whether those actions were wrongful (e.g. Epstein 1973). Others have argued that a court should be allowed to hold a defendant liable for a claimant's losses even in the absence of evidence that the defendant's actions contributed to a causing of the harm, especially in so-called 'mass tort' cases where the causal facts may be complex and/or unclear (e.g. Strudler 1992; though see Schroeder 1990; Steel 2012). It has been argued, on these and other grounds, that the best way to understand tort law is not in conceptual, deontological terms, but rather in instrumentalist, utilitarian terms (Summers 1982; Posner and Landes 1980). I won't get into these controversies here. Rather, I want to advance a novel critique of (WC), one which focuses on its failure to recognise the possibility of *partial* liability.

3. Proximate Causation and the Relational Account of Wrongfulness

In the classic *Palsgraf v. Long Island Railroad Co.*,[2] an employee of the defendant company negligently helped a man board a moving train. The man lost his balance and dropped a package containing fireworks on to the platform. It exploded on impact, sending shockwaves to the other end of the platform which tipped over a set of scales onto one Mrs. Palsgraf, injuring her. Palsgraf's injury wouldn't have occurred but for the negligence of the train guard. Yet the question of liability is less clear – intuitively, the injury seems in some sense *too removed* from the initial wrongdoing for it to be reasonable to hold the train company (as the guard's employer) liable for it.

In practice, the common law handles these kinds of cases by invoking the legal concepts of 'proximate causation' or 'remoteness'.[3] Although the action of the train guard in *Palsgraf* was a 'factual cause' of Palsgraf's injury, the train company is liable for Palsgraf's losses only if the action of the train guard was a 'proximate cause' of the injury. Legal theorists have spent a lot of time worrying about this distinction between 'proximate' and 'factual' causation.[4] The current consensus seems to be that "'proximate cause' is neither about cause nor proximity, as those two words are commonly understood".[5] It can't be about *spatiotemporal* proximity, because defendants are rightly held liable for losses arising from harms greatly removed both in space and time from their wrongdoing. In *People v. Botkin*,[6] for example, the defendant sent a box of poisoned candy from San Francisco to Dover, in Delaware. The

[2] 248 N.Y. 339 (1928).

[3] 'Proximate causation' is more common in American jurisdictions, with England generally preferring 'remoteness'. I'll continue to use the former term here.

[4] See Beale (1920); Edgarton (1924); Green (1927); Hart and Honoré (1985); James and Perry (1951); Malone (1956); Stapleton (2001); Wright (1985).

[5] American Law Institute, *Restatement (Third) of Torts: Liability for Physical and Emotional Harm*, §26 cmt. a (2009).

[6] 132 Cal. 231 (1901).

victim ate the candy and died. Whether the defendant is liable for the claimant's losses in *this* case clearly has nothing to do with the distance between San Francisco and Deleware or the time it took for the candy to arrive. But 'proximate causation' can't be interpreted in *causal* terms either, according to the most recent Restatement of Torts, because "[a] necessary condition for a relevant harm is a factual cause of that harm, without limitation" and "there are no degrees of factual cause".[7] Once it's been established that the defendant's wrongdoing was a cause of the claimant's harm, it's generally assumed that any further questions couldn't possibly belong to any genuine causal inquiry.

We will, of course, return to this latter assumption below. But it will help in motivating my alternative account of proximate causation to examine the most popular attempt to make sense of it within the framework of the corrective justice approach – the so-called *relational account of wrongfulness*. It received its canonical formulation in Chief Justice Cardozo's majority opinion to *Palsgraf*. Cardozo agreed that, since the train guard's action was clearly a cause of Palsgraf's injury, the controversy over whether to hold the train company liable for Palsgraf's losses had nothing to do with either proximity or causation: "The law of causation, remote or proximate, is thus foreign to the case before us".[8] He also agreed that the train guard bore a duty to Palsgraf, as well as to everyone else in the station, not to wrongfully contribute to a causing of harm to her. After all, as Justice Andrews notes in his dissent, "[d]ue of care is a duty imposed on each one of us to protect society from unnecessary danger, not to protect A, B or C alone".[9] Nevertheless, Cardozo insisted that the train company shouldn't be held liable for Paslgraf's losses. An action isn't simply wrongful *simpliciter*, he argued, but only *relative* to an outcome: "Negligence in the air, so to

[7] *Restatement*, *supra* note 4, §26 cmt. j.

[8] *Palsgraf*, *supra* note 2, at 346.

[9] *Id.*, at 349.

speak, will not do", to quote Pollock's (1920: 455) well-known maxim. To wrongfully contribute to a causing of harm, one must perform an action which contributes to a causing of a harm, and that action must be wrongful *relative to that harm*. It's this latter condition, according to Cardozo, that isn't satisfied in *Palsgraf*. Although the action of the train guard was wrongful relative to any harms the holder of the package may have suffered, Cardozo argued that it wasn't wrongful relative to Palsgraf's injury, and so the train guard didn't breach his duty to Palsgraf not to wrongfully contribute to a causing of harm to her. "The conduct of the defendant's guard, if a wrong in its relation to the holder of the package, was not a wrong in its relation to the plaintiff, standing far away. Relatively to her it was not negligence at all."[10]

What is it for an action to be wrongful *relative* to a particular harm? The most common approach here appeals to the concept of *reasonable foreseeability*. According to Lord Simmons, for example, "it would be wrong that a man should be held liable for damage unpredictable by a reasonable man…[t]hus foreseeability becomes the effective test".[11] In other words:

> TOKEN FORESEEABILITY (TOK-F): S's φ-ing at *t* was wrongful relative to a harm *H* only if *H* was reasonably foreseeable by S at *t*.

According to (TOK-F), the train guard's actions weren't wrongful relative to Palsgraf's injuries, because those injuries weren't reasonably foreseeable by the train guard when he helped the man board the moving train. So (TOK-F) correctly predicts that Palsgraf's claim should fail, on the grounds that the train guard did not breach the duty he bore to Palsgraf not to wrongfully contribute to a causing of harm to her.

---

[10] *Id.*, at 341.

[11] *Overseas Tankship (UK) Ltd v. Morts Dock & Engineering Co. Ltd (The Wagon Mound (No. 1))*, [1961] AV 388, 426 (JCPC).

As Hart and Honoré (1985) have argued, however, (TOK-F) is actually too strong. Suppose Zach shoots at Yvonne, intending to kill her. At the last minute Yvonne ducks, and the bullet hits and kills Walter who, unbeknownst to Zach, was standing behind Yvonne. Walter's death wasn't reasonably foreseeable by Zach at the time he shot his weapon – no reasonable person in Zach's epistemic position would have foreseen Yvonne's ducking at just the right time and Walter's standing in just the wrong place.[12] But the law in such cases has typically held that Zach ought nevertheless to be held liable for any losses arising from Walter's death:[13] When a defendant "ought to have foreseen in a general way consequences *of a certain kind*, it will not avail him to say that he could not foresee the precise course or the full extent of the consequences, being of that kind, which in fact have happened" (Pollock 1922: 167; my emphasis). Although Zach couldn't reasonably have foreseen the *particular* harm that in fact occurred, it seems he should nevertheless be held liable for it, because his actions were intended to bring about a harm of the very kind which in fact materialised.

This suggests a weaker condition on relational wrongfulness:

> TYPE FORESEEABILITY (TYP-F): S's φ-ing at *t* was wrongful relative to a harm *H* only if it was reasonably foreseeable by S at *t* that a harm of type **H**, of which *H* was an instance, would occur if she φ-ed.

[12] I'm implicitly assuming here that the subject of a death is essential to it – the death that in fact befalls Walter wouldn't have occurred had the bullet hit Yvonne instead. But a counterpart-theorist believes that such claims are relativized to a choice of counterpart relation (see Chapter Three, above); and she may insist that there is *some* counterpart relation relative to which Walter's death would still have occurred had Yvonne not ducked (only it would have been a death of Yvonne, not a death of Walter). Relative to this counterpart relation, Walter's death *was* reasonably foreseeable by Zach. The challenge for the counterpart-theorist then becomes that of specifying *which* counterpart relation is the right one to use for the purposes of determining an agent's culpability; and many of the issues discussed in relation to (TYP-F) below will come up for the counterpart-theorist too, albeit under a different guise.

[13] C.f. *Bici v. Ministry of Defence* [2004] EWHC 786, which confirms English law's commitment to a version of the 'transferred malice' doctrine for the tort of battery.

Foreseeing *that* a harm of a certain kind will occur is not the same as foreseeing a particular harm.[14] In the case above, for example, Zach could reasonably have foreseen *that* a death would occur if he fired his weapon, even though he couldn't reasonably have foreseen the death that in fact occurred. Thus (TYP-F) correctly implies that Zach should be held liable for Walter's death.[15]

A number of theorists have explicitly expressed the hope that the relational account of wrongfulness, understood along the lines of (TYP-F), could completely supplant the language of 'proximate causation' in the law (e.g. Seavey 1939). On this view, all 'proximate cause' problems boil down to whether there is a mismatch "between the actual result and the [type of] result intended or risked" (Kadish *et al.* 2001: 517). I am less optimistic. (TYP-F) faces some serious challenges; and even if those challenges can be met, there is reason to think that the relational account of wrongfulness couldn't in principle do away with the need for a proximate causation requirement.

The first challenge for the proponent of (TYP-F) is to specify restrictions on allowable types of harm. Without such restrictions, (TYP-F) is simply vacuous. Palsgraf's injury was an instance of the type of harm *injury to Palsgraf caused by falling scales*, and the train guard certainly couldn't have foreseen that a harm of *that* type would have resulted from helping the man board the moving train. But Palsgraf's injury was also an instance of the more general type of harm, *injury*; and the train guard certainly *could* have foreseen that harm of *that* type might have resulted from helping the man board the train. So for (TYP-F) to have any substantive content, we should understand

[14] C.f. the distinction between *seeing that a table exists* and *seeing a table*; I can see that a table exists without seeing a table (I might read a newspaper article about tables, for example), and I can see a table without seeing that a table exists (I might lack the concept of a table). See Dretske (1969).

[15] It's worth pointing out, however, that there may be Gettier-style obstacles to Zach's foreseeing that a death will occur in such a case. After all, if Zach believed that a death would occur if he shot his weapon, there's a sense in which that belief was based on a false premise, namely that it is *Yvonne* that would die. For arguments that foreseeing is a kind of knowing, see Williamson (2000).

it as restricted to sufficiently *fine-grained* types of harm. On the other hand, the types of harm shouldn't be *too* fine-grained either. Suppose Malcom answers his mobile phone whilst driving. His action poses a small risk of a rear-end collision occurring, a small risk of a collision with a pedestrian occurring, a small risk of damage to a home-owner's garden occurring, and so on, for many different types of harm. But it might well be that Malcolm couldn't reasonably have foreseen, for any *one* of these types, that a harm of that type would occur. If types of harm were restricted to these fine-grained classifications, then, (TYP-F) would imply that Malcolm shouldn't be held liable for any harms that materialize as a result of his actions, which seems like the wrong result; as Hurd and Moore point out, defendants in such cases "patently know or should know that they have no business doing what they are doing" (Hurd and Moore 2002: 383).

The law itself has had little to say on this matter, except that "the inquiry…into the nature of the risks or hazards, the foreseeability of which makes conduct negligent, must be neither too refined nor too coarse" (James and Perry 1951: 799). The challenge for the proponent of (TYP-F) is to show that there is a partition of the set of harms which satisfies both of these criteria. But even if this challenge could be met, the relational account of wrongfulness still can't entirely do away with the need for a proximate causation requirement. To see this, consider the following case. Suppose Hussein recklessly runs a red light and collides with Idris, who suffers a broken arm in the collision. As a result, Idris doesn't move house that month as planned. Many years later, he suffers another broken arm in another car accident at the same junction. Neither broken arm would have occurred but for Hussein's running the red light, since if he hadn't run the red light, Idris would have been many miles away from the junction at the time of the second accident. But Hussein should clearly only be held liable for the losses arising from the first injury, and not those arising from the second. The problem is that it's not at all obvious how the relational account of wrongfulness can

explain this. After all, *both* injuries were breakings of Idris's arm in traffic collisions.[16] However we carve up the set of harms, then, if Hussein is liable for the losses arising from the first injury in virtue of it being reasonably foreseeable by him that a harm of a certain kind, of which Idris's injury was an instance, would occur if he ran the red light, (TYP-F) doesn't rule out holding Hussein liable for the losses arising from the second injury too, *for the very same reason*. The problem is that (TYP-F) "only asks after a logical relation between a type of harm (the one the risk of which made the defendant negligent) and a token of harm (the harm that actually happened)", and this logical relation "takes no notice" of the "freakishness" of the "causal route" (Hurd and Moore 2002: 405). It's precisely this notion of the 'freakishness' of a causal route that the proximate causation requirement seems designed to capture.

In light of these challenges, some writers ultimately deny that the proximate causation requirement can, or indeed should, be explained in corrective justice terms. Stapleton, for example, describes the "obfuscating terminology of…proximate cause" (Stapleton 2001: 945) as a mere collection of "pseudo-scientific metaphors [which] serve only to mask the nature of the underlying dispute about competing visions of the scope of liability" (Stapleton 2001: 969), a dispute which primarily turns on questions of legal policy (c.f. Chapter One, Section 5, above). Accordingly, although the Restatement (Third) of Torts reluctantly acknowledges the widespread usage of the term by including 'Proximate Cause' in parentheses in its section on 'Scope of Liability', it also "fervently hopes that the Restatement Fourth of Torts will not find this parenthetical necessary".[17]

[16] Of course, one *could* insist that Hussein's actions were wrongful relative to breakings of Idris's arm *within a minute of Hussein's action*, but not wrongful relative to breakings of Idris's arm *within a year of Hussein's action*. But this move is clearly *ad hoc*, at least without an independent explanation of why temporal separation should matter.

[17] *Restatement*, *supra* note 4, special note on proximate cause.

Despite repeated attempts to eradicate it, however, the language of proximate causation remains as stubbornly pervasive as ever in legal discourse. I think there is a good reason for this. Although talk of 'proximate causation' cannot be replaced by a relational account of wrongfulness, this doesn't mean that we can't make sense of it within the corrective justice framework. Cardozo assumed that the concept of proximate causation couldn't be interpreted in causal terms, because the only relevant causal question it makes sense to ask is whether the defendant's wrongdoing was a cause of the claimant's harm. But this assumption is mistaken. Once we've established that the defendant's wrongdoing was a cause of the claimant's harm, there is always a further question of the *degree* of contribution the wrongdoing made to the causing of the harm. In the next section, I will defend a causal theory of proximate causation in terms of this notion of degrees of causal contribution.

4. A Causal Theory of Proximate Causation

In Chapter Five, I argued that causation, though not a scalar relation, is nevertheless a relation to which different events can *contribute* to different degrees. I therefore propose to interpret tort law's proximate causation requirement as expressing a commitment to the following characterization of the duties of corrective justice:

THRESHOLD (T): Where D and P are legal persons:

- D bears a first-order duty to P not to wrongfully contribute to a causing of harm to P *to a significant degree*, and
- if D breaches her first-order duty to P, D acquires a second-order duty to restore P to the state she would have been in, but for the harm.

In other words, a 'proximate cause' should be thought of as a cause which contributed to a causing of an effect to a significant degree – a degree above a certain threshold.[18]

In Chapter Five, I also defended the following account of causal contribution:

> CAUSAL CONTRIBUTION (CC): If $C_1,\ldots, C_n$ collectively caused E relative to **B** and **O**, then $f(C_i, [C_1,\ldots,C_n]\rightarrow E)^{\mathbf{B},\mathbf{O}} = \dfrac{P(E \mid C_i \wedge \max(\mathbf{B}, \mathbf{O}))}{\sum_{j=1}^{n} P(E \mid C_j \wedge \max(\mathbf{B}, \mathbf{O}))}$

(T) and (CC) can explain the decision in *Palsgraf*. The important fact is that, even given that the man dropped the package of fireworks, the chance of the shockwaves following exactly the trajectories required for tipping the scales in exactly the way required for Palsgraf's injury to occur, is vanishingly small – the vast majority of relevant possibilities are ones in which Palsgraf's injury doesn't occur. Hence, at least relative to a context in which the defendant's wrongdoing was one of a number of causes of Palsgraf's injury, it follows from (CC) that the wrongdoing's degree of contribution to the causing of the injury is negligible. *That's* why it seems unreasonable to hold the train company liable for Palsgraf's losses – the train guard didn't breach the duty he bore to Palsgraf, according to (T), because this duty was a duty not to wrongfully contribute *to a significant degree* to a causing of harm, and his actions didn't contribute to a significant degree to the causing of Palsgraf's injury.

(T) and (CC) can also explain many of the 'tests' for proximate causation that judges have developed. In his dissent in Palsgraf, for example, Andrews lists "some hints that may help us" in determining whether one event was a 'proximate cause' of another:

> Was the one a substantial factor in producing the other? Was there a direct connection between them, without too many intervening causes? Is the effect of cause on result not

[18] What counts as a 'significant' degree of contribution is a vague matter, of course, but an element of vagueness in the law should come as no surprise.

> too attentuated [sic]? Is the cause likely, in the usual judgment of mankind, to produce the result?[19]

All of these questions make sense in light of (T) and (CC). Plausibly, what it means to say that an event was a 'substantial factor'[20] in producing an effect (or that it made a 'material contribution', to use the phrase more common in England[21]), is just that it contributed to a significant degree to a causing of the effect. Talk of 'the effect of cause on result' as 'not too attenuated' can be interpreted in much the same way. It also makes sense to ask how many other events were involved in the causing of the effect – as discussed in Chapter Five, it follows from (CC) that the degree of contribution of an event to a causing of an effect decreases as the number of other causes increases, all other things being equal. Finally, it follows straightforwardly from (CC) that the 'likelihood' of the effect occurring given that the cause occurred is a good guide to whether the event made a significant contribution to the causing of the effect.

Moore (2009) has recently defended a similar principle to (T): "[T]he amount of causal contribution needed for an actor to be morally responsible for some harm is non-*de minimis* (or 'substantial')." Unfortunately, however, Moore doesn't explain what 'causal contribution' is, except to say that it "peters out over time, much as the ripples from a stone dropped in a pond diminish as they travel outward" (Moore 2009: 276); a metaphor which suggests that a defendant shouldn't be held liable for losses arising from harms which occur a long time after the wrongdoing, which as we've seen (and as Moore himself accepts) gets the wrong result in some cases. It doesn't follow from (CC), however, that 'proximate' causes need be *spatiotemporally* proximate to their effects. In *Palsgraf*, for instance, the injury occurred fairly close to, and fairly soon after, the wrongdoing. The relevant factor is the probability of the injury occurring

---

[19] *Palsgraf*, *supra* note 2, at 354.

[20] See Smith (1911).

[21] This terminology appears to have originated in *Duke of Buccleuch v. Cowan* (1866) 5 M. 214; see Steel and Ibbetson (2011).

conditional on the wrongdoing occurring, not the spatiotemporal separation of the two events.

That being said, however, (CC) can also explain why talk of 'proximity' wasn't entirely off the mark in the first place. After all, spatiotemporal proximity is undoubtedly a decent *guide* to liability in many cases, as even Cardozo admits: "There is no use arguing that distance ought not to count, if life and experience tell us that it does".[22] The reason for this, on my view, is that many systems in the natural world are approximately *chaotic*. What this means, in essence, is that the chance of the system being in any particular specific macroscopic state M at t, conditional on its being in state M_0 at an earlier time t_0, exponentially tends to zero as $t - t_0$ tends to infinity (the package of fireworks is plausibly such a system – arbitrarily small changes in the positions of the fireworks in the package at one time can lead to large differences in where the shockwaves travel when the package explodes). In chaotic systems, therefore, degrees of causal contribution do indeed decrease with temporal separation between the cause and the effect, all other things being equal. But many systems are not chaotic, as evidenced by the mere possibility of macroscopic predictive science. In these systems, the degree of contribution of one event to a causing of another need not depend in any systematic way on the time interval between them.

As explained in Chapter Five, an event contributes to a causing of an effect only *relative* to a modal base and ordering source. An inquiry into whether a defendant is liable for a claimant's losses may therefore involve two distinct tasks. The first is the meta-task of determining relative to *which* modal base and ordering source (T) is true; the task, in other words, of selecting the modal base and ordering source relative to which causal claims ought to be evaluated for the purposes of establishing whether an agent has breached a legal duty. This may well involve considerations of legal policy,

[22] *Bird v. St. Paul F. and Minneapolis Ins. Co.* 224 N.Y. 47 (1918).

as well as reflections on our concept of moral responsibility (c.f. Chapter Two, Section 6, above). Once the right contextual parameters are fixed, however, the second task – that of establishing the degree of contribution of the defendant's wrongdoing's to the causing of the claimant's harm – is a purely causal inquiry.

Although the language of 'proximate causation' is used in a variety of different ways for different purposes in the law, this section has argued that there is a concept of proximate causation *worthy of the name*; that certain legal doctrines are successfully tracking this concept; and that this concept is best interpreted in terms of (T). In the next section, however, I will argue that proximate causation doctrines ought after all to be replaced with a more fine-grained approach, one which recognises the possibility of *partial* liability for losses.

5. In Defence of Partial Liability

(T), just like (WC), implies that liability is all-or-nothing. A defendant either breached a duty to the claimant not to wrongfully contribute to a significant degree to a causing of harm to him, or she didn't. If she did, she is bound by a second-order duty to restore the claimant to the state he would have been in but for the harm, and so is liable for 100% of the claimant's losses; if she didn't, she bears no second-order duty to the claimant at all. But this all-or-nothing approach has increasingly come to be viewed as too inflexible in cases where the claimant's harm was collectively caused by the wrongdoing of multiple individuals. These can be divided into two categories, usually treated separately. The first kind of case is one in which the claimant's negligence contributed to the causing of his own harm. Here's an example:

> *Fire 4*
>
> Big Corp negligently fails to replace its outdated equipment. The equipment short circuits. At around the same time, Greta negligently drops a lighted cigarette nearby. Both events start small fires which, if either had occurred alone, would eventually have fizzled out

> without causing much damage. Unfortunately, the two fires join together into one big fire, which burns down Greta's house.

According to (T), Big Corp in *Fire 4* is either fully liable or not at all liable for Greta's losses, depending on whether Big Corp's negligence made a significant contribution to the causing of the damage. But this all-or-nothing approach has recently been "roundly criticised for its harshness and unfairness" (Gardner 1996: 3). Requiring Big Corp to absorb *all* of Greta's losses seems too harsh, since, after all, the damage wouldn't have occurred but for Greta's own negligence. But excusing Big Corp of *all* liability seems too lenient, since, after all, the damage wouldn't have occurred but for Big Corp's negligence.

A number of legislative overhauls have sought to allow for more flexibility in such cases. The paradigm example is the *Law Reform (Contributory Negligence) Act 1945*, which states that:

> Where any person suffers damage as the result partly of his own fault and partly of the fault of any other person or persons…the damages recoverable in respect thereof shall be reduced to such extent as the court thinks just and equitable having regard to the claimant's share in the responsibility for the damage.[23]

The act allows a court to reduce the damages owed to the claimant, in principle by any fraction of the claimant's total losses, so as to reflect the claimant's own 'share in the responsibility' for the harm caused. Similar reforms have since been introduced in the majority of Anglo-American jurisdictions.[24]

---

[23] *Law Reform (Contributory Negligence) Act 1945*, 8 & 9 Geo. 6 c. 28, §1, sch 1.

[24] In the US, so-called 'comparative negligence' rules have been gradually introduced by statute; see American Law Institute, *Restatement (Third) of Torts: Apportionment of Liability* (2000). Exceptions (as of writing) include Alabama, Maryland, North Carolina, Virginia, and the District of Columbia, where contributory negligence remains a complete defence, defeating all liability for the claimant's losses.

The second kind of case is one in which the claimant's harm was collectively caused by the negligence of multiple defendants. Here's an example:

> *Fire 5*
>
> Big Corp and Little Corp both negligently fail to replace their outdated equipment. Both sets of equipment short circuit. Both events start small fires which, if either had occurred alone, would eventually have fizzled out without causing much damage. Unfortunately, the two fires join together into one big fire, which burns down Greta's house.

According to (T), Big Corp and Little Corp in *Fire 5* are both fully liable for 100% of Greta's losses, so long as their actions each made a significant contribution to the causing of the damage. This is reflected in traditional 'joint and several liability' systems, under which Greta would be free to choose to sue either company for all her losses.[25] These systems work well for claimants, of course, who are spared the often costly and time-consuming task of tracking down every relevant tortfeasor. But defendants have long argued that this system places an unreasonable burden on those with the deepest pockets, who inevitably end up being the ones footing the bill, even in cases where their actions contributed very little to the causing of the harm. In *Fire 5*, for example, Greta is likely to sue Big Corp rather than Little Corp, regardless of the degree of contribution of Big Corp's negligence to the causing of the damage, simply because Big Corp, being bigger, is less likely to default on its second-order duty. Critics of joint and several liability systems have had some limited but notable successes (see Barker and Steele (2015) and references therein). The clearest example is the so-called 'proportionate liability'[26] systems introduced in Australia in the early 2000s, which

[25] At common law, if D1's negligence and D2's negligence collectively causes V's harm, and V sues D1 for all her losses, D2 wouldn't have to pay anything to anyone. Most jurisdictions have since introduced – allegedly on grounds of 'fairness' – statutory mechanisms allowing D1 to bring a subsequent claim against D2 for a contribution towards the damages paid to V in such cases (see Rogers and Boom 2004). As discussed below, however, it's difficult to justify these doctrines on corrective justice grounds, since D1 plausibly hasn't been *wronged* by D2 just because V chose to sue the latter.

[26] Note that what I mean by 'proportionate liability' here is very different from what might be called 'probabilistic liability', where liability is apportioned according to the *epistemic* probability that each defendant's actions was a cause of the claimant's harm (see Steel 2015:

require that "the liability of a defendant…is limited to an amount reflecting that proportion of the damage or loss claimed that the court considers just having regard to the extent of the defendant's responsibility for the damage or loss".[27]

There is a natural way of interpreting these reforms, in light of the resources developed above. Lawyers and legislators are gradually coming to realise that (T) doesn't capture the essence of corrective justice. Once we realise that contributions to causings come in degrees, we should *also* think of liability as coming in degrees. In other words, (T) should be replaced with the following principle:

> PROPORTIONALITY (P): Where D and P are legal persons:
>
> – D bears a first-order duty to P not to wrongfully contribute to a causing of harm to P (to *any* degree), and
> – if D breaches her first-order duty to P by wrongfully contributing to a causing of harm to P *to degree x*, D acquires a second-order duty to *contribute to degree x* towards restoring P to the state she would have been in, but for the harm.

(P) encodes a natural relationship between the contribution a defendant's wrongdoing makes to the causing of a harm and the contribution she is thereby required to make towards repairing it. Suppose, for example, that Big Corp contributes to the causing of Greta's harm to degree 0.4 in *Fire 5* (so that Little Corp contributes to degree 0.6, since degrees of contribution to causings always sum to 1). Then Big Corp acquires a second-order duty to *contribute to degree 0.4* towards restoring Greta to the state she would have been in but for the damage to her house. Big Corp can discharge this duty by

ch. 6). My view has nothing to say about how to proceed in cases where the causal facts are uncertain.

[27] *Civil Liability Act 2002* (NSW) s 35 sch 1 part a. Most American states have also introduced proportionate liability for some kinds of harm, retaining traditional joint and several liability for others; see *Restatement*, *supra* note 24. These reforms have been controversial, however, and some jurisdictions remain firmly opposed to them; see Barker and Steele (2015); Wright (1988); McNichols (1979).

paying monetary damages to Greta equal to 40% of Greta's total losses. I'll say that Big Corp is *liable to degree 0.4*, or 40% liable, for Greta's losses in such a case. Note, however, that being 40% liable for a claimant's losses is *not* the same as being liable for 40% of the losses. D is *liable for 40%* of P's losses just in case D's wrongful conduct caused a harm to which 40% of P's total losses are directly attributable; D is *40% liable* for P's losses just in case D's wrongful conduct contributed to degree 0.4 to a causing of a harm to which 100% of P's total losses are directly attributable. These are different second-order duties, even if they can be discharged in the same way.

Now consider yet another variant on the *Fire* case:

> *Fire 6*
>
> Big Corp and Little Corp both negligently fail to replace their outdated equipment. Both sets of equipment short circuit at the same time. Both events start large fires which spread independently and arrive at Greta's house at the same time. Greta's house is badly damaged.

Under joint and several liability, *Fire 6* would be resolved in exactly the same way as *Fire 5* – Greta would have the option of suing either Big Corp or Little Corp for 100% of her losses, at which point the company she chose would then have a claim against the other company for a contribution towards the damages paid.[28] But the two cases are treated very differently by (P).[29] Whereas in *Fire 5* Big Corp's negligence and Little Corp's negligence *collectively* caused the damage, in *Fire 6* they each *individually* caused the damage. Big Corp and Little Corp are therefore both fully liable (i.e. liable

[28] E.g. the law in England states that "any person liable in respect of any damage suffered by another person may recover contribution from any other person liable in respect of the same damage (*whether jointly with him or otherwise*)" (*Civil Liability (Contribution) Act 1978*, s1(1), my emphasis).

[29] Thus joint and several liability systems and those based on (P) don't simply differ in their assessments of who "should bear the expense of apportionment and the risk of collecting from insolvent or otherwise unavailable tortfeasors" (Wright 1988: 1143) – (P) also draws important distinctions between cases of joint causation and cases of overdetermination that are not recognised under joint and several liability.

to degree 1) for the damage, according to (P). Hence in *this* case, Greta *can* choose to sue either company for 100% of her losses. Suppose she chooses Big Corp; what then happens to Little Corp? Little Corp, just like Big Corp, breached a first-order duty which gave rise to a second-order duty to restore Greta to the state she would have been in, but for the damage to her house. But once Big Corp pays up, Greta is *already* in the state she would have been in, but for the damage to her house. Hence Little Corp is not required to make any contribution towards Greta's losses, according to (P) – after Greta receives damages from Big Corp, Little Corp's second-order duty of repair evaporates, on account of there being nothing left to repair. Moreover, Big Corp now has no right against Little Corp to recover any of the damages it paid to Greta – after all, Big Corp has not been *wronged* by Little Corp, and so there is no wrong such a settlement could be interpreted as correcting. (If this seems unfair to you, do not be alarmed; your concerns will be addressed below.)

In Chapter Five, I argued that an event can contribute to different degrees to different causings of the very same effect. In *Committee*, for example, D_2's vote contributed to degree 0.42 to one causing of the employee's injury and it contributed to degree ⅓ to a different causing of the injury. To what degree, then, should D_2 be held liable for the employee's losses in such a case? On the face of it, there are a number of options. One option would be to add the two degrees of contribution together, so that D_2's degree of liability for the employee's losses is (about) 0.75. Another option would be to take the average of the two values, so that D_2's degree of liability for the employee's losses is (about) 0.38. I think neither of these strategies is ultimately correct. Instead, D_2 should be held liable for the employee's injury to degree 0.42 – the larger of the two values.

To see this, suppose first that Nina poisons Oscar's lunch. But she wants to make sure of his death – so at just the moment when Oscar is about to succumb to the poison, she fatally shoots him. Oscar's death is overdetermined by the shooting and the poisoning. In this case, then, Nina performed two actions, each of which individually

caused a death. But it surely wouldn't be right to charge Nina with *two* counts of murder. To do so would intuitively involve some objectionable double-counting, since it was *the same death* caused twice over.

Now consider a slightly different case. As before, Nina poisons Oscar's lunch, and also shoots him at just the time he is about to succumb to the poison. But this time, her gunshot wounds Oscar non-fatally. However, at the same time, Patrick also shoots Oscar non-fatally, so that the two gunshots collectively necessitated Oscar's death. As before, Oscar's death was caused twice – once by the poisoning and once by the two gunshot wounds. Suppose Nina's gunshot contributed to degree 0.5 to the latter causing. Hence in this case, Nina contributed to degree 1 to one causing of Oscar's death and also contributed to degree 0.5 to a different causing of Oscar's death.

I think the right reaction in this case is again to charge Nina with a single count of murder. To punish her separately for her contributions to *both* causings of Oscar's death would involve the same kind of objectionable double-counting as in the case above. But on the other hand, it would be absurd to reason that, since her *average* contribution to causings of Oscar's death was 0.75, she should be punished *less* than if she had merely murdered Oscar by poisoning alone.

These cases aren't exactly analogous to the *Committee* case, because Nina in both cases performs two *separate* actions, each of which contributes to a different causing of Oscar's death; whereas in *Committee*, D_2 performs a *single* action which *itself* contributes to two different causings of the employee's injury. Nevertheless, I think, similar considerations apply. To hold D_2 liable to a degree equal to the *sum* of the two degrees of contribution would involve some objectionable double-counting, because it's the *same injury* being caused twice over. To take the *average* of the two degrees of contribution, on the other hand, would be to let D_2 off too lightly. So we should take the larger of the two degrees of contribution, and ignore the other. This means that D_2,

D_3 and D_4 should each be held liable to degree 0.42 for the employee's losses, and D_1 should be held liable to degree 0.58. And *this* means the employee could sue D_2, D_3 and D_4 for *up to* 42% of his losses and D_1 for *up to* 58% of his losses, until such time as his losses are fully compensated; at which point any remaining second-duties of repair evaporate, on account of there being nothing left to repair.

(P) is, I hope, fairly intuitive on its face. It's simply the principle that defendants should be held liable for losses arising from harms only to the extent to which their wrongdoing contributed to bringing those harms about.[30] It's also related to (T) in a natural way: (T) can be thought of as a *coarse-grained* version of (P), one which follows from imposing on (P) the unmotivated constraint that liability must be all-or-nothing. Whereas according to (T), a defendant's degree of liability effectively jumps from 100% to 0% as the degree of contribution of her wrongdoing to the causing of the harm passes the threshold, (P) allows for degrees of liability to gradually decrease as a function of the wrongdoing's degree of contribution to the causing of the harm.

6. Objections and Replies

There are two main objections to partial liability in the literature. The first is that it cannot provide claimants with adequate compensation in cases where one or more of the defendants are insolvent or untraceable. As a number of commentators have pointed out, compared with joint and several liability, (P) seems strongly weighted in favour of defendants.[31] Consider *Fire 5* again, in which the damage to Greta's house was collectively caused by Big Corp's and Little Corp's negligence. Suppose Little Corp has no liability insurance and no assets, or it's currently insolvent, or it no longer exists, or for some other reason it isn't able to discharge its second-order duty to Greta. Under joint and several liability, Greta can still recover 100% of her losses from Big

[30] Interestingly, a number of authors have defended similar principles for moral responsibility; see Bernstein (forthcoming) and Sartorio (forthcoming).

[31] See Barker and Steele (2015); Wright (1988); McNichols (1979); etc.

Corp. But under (P), Greta can only sue Big Corp for 40% of her losses, and must absorb the remaining 60% herself. This seems unfair; after all, Big Corp is a large multinational corporation with a poor safety record for which 60% of Greta's losses is mere small change, whereas Greta is an innocent victim who has lost everything she owns in a fire she neither caused nor could have prevented. It feels morally unacceptable to ask Greta to absorb even a fraction of her losses, given that her losses could so easily be absorbed by Big Corp (or, more likely, by Big Corp's liability insurance provider).

I agree that there is *something* unjust about Greta having to absorb 60% of her losses in *Fire 5*. But this is a *distributive* injustice, not corrective one. To illustrate, consider *Fire 2* again. In this case, Big Corp negligently failed to replace its outdated equipment, but the equipment didn't short circuit. Instead the damage to Greta's house was caused by a lightning strike. Although Big Corp's negligence *could* easily have started a fire, it actually played no role at all in bringing about the damage. Both (P) and (T) (and, for that matter, (WC) too) imply that Greta cannot recover her losses from Big Corp in this case. From a *distributive* perspective, it might seem unfair that Greta must absorb her losses when they could be so easily absorbed by less deserving individuals, Big Corp included. But the duty to compensate Greta in *Fire 2*, if there is one at all, falls not to Big Corp but to the State, or failing that, Greta's community at large. If Greta is owed compensation in *Fire 2*, it is not because she has been *wronged*, but because one has a duty to distribute (and *re*distribute) resources in such a way as to alleviate the suffering of those harmed through no fault of their own. The fact that Greta's harm can be alleviated at such a small cost to Big Corp is, of course, not irrelevant to the question of how to meet the demands of distributive justice in this case, but it is of no relevance to the demands of corrective justice.

I think that tort law is a system of corrective justice. I *don't* think that it is, even in part, a system of distributive justice.[32] The demands of tort law might *coincide* with the demands of distributive justice in some cases, but this is merely a side-effect. There are other systems, such as taxpayer-funded state welfare, which play the role of discharging the demands of distributive justice already (or at any rate, there ought to be).[33] Thus I don't think that it is part of the remit of tort law to provide compensation to Greta in *Fire 2*, no matter how deserving she might be. This much is fairly uncontroversial – every system of tort law, as far as I know, would require Greta to prove that there was some causal connection between Big Corp's negligence and the damage to her house in order to successfully claim damages from Big Corp in *Fire 2*. But exactly the same considerations apply to *Fire 5*. Big Corp has wrongfully contributed to the causing of the damage to Greta's house to degree 0.4. This is a wrong for which contributing to degree 0.4 towards compensating Greta for her losses is the required remedy. Once Big Corp has discharged this duty, the demands of *corrective* justice have been met. How Greta ought to be compensated for the remainder of her losses, if at all, is an important question, but it's a question of distributive justice, not corrective justice, and hence not a question to which tort law need provide an answer.

Some might insist that, even if it's fundamentally a system of corrective justice, tort law can still legitimately be used to pursue distributive goals, perhaps in order to pick up some of the slack created by the state's failure to address them.[34] But the danger with this approach is that we end up with a tort system which cannot be justified on

---

[32] Thus I disagree with Part 2 of Gardner's (2014) analysis of 'what tort is law is for'. Gardner claims that apportionment of liability mechanisms in cases with multiple defendants "lack a corrective-justice rationale", since "[c]orrective justice…knows only addition and subtraction", and hence "has no room for division, which is the business of distributive justice" (Gardner 2014: 349). As explained above, I think that apportionment of liability mechanisms *can* be explained in purely corrective terms, using the concept of partial liability.

[33] Some countries also have universal no-fault compensation schemes for certain kinds of harms – see the *Accident Compensation Act 2001* (N.Z.).

[34] Coleman has consistently argued that "tort law is best explained by corrective justice" (Coleman 1992: 9), but even he concedes that there are "other goals the state may legitimately pursue within the tort system" (Coleman 2001: 392).

either corrective or distributive grounds. Joint and several liability systems would hold Big Corp 100% liable for the damage to Greta's house in *Fire 5*. 40% of this liability can be justified on corrective grounds, on my view, so the other 60% must be justified on distributive grounds. We might appeal to the fact that Greta is deserving and poor, whereas Big Corp is undeserving and rich, for example. But there are many deserving, poor people besides Greta – why should Big Corp pay extra money to *Greta* rather than these other people? There are also many undeserving, rich people, and many more undeserving, rich corporations – why should *Big Corp*, rather than these other individuals, cover the remainder of Greta's losses? It's no use pointing to the causal connection between Big Corp's negligence and Greta's harm, because this fact is simply irrelevant to the question of how to fairly distribute the costs of the damage.

For these and other reasons, I think that normative coherence in tort law can only be achieved by using it purely to pursue the demands of corrective justice, and this means replacing (T) with (P) across the board. Holding everything else fixed, such a change would probably increase the level of distributive injustice in the world, since it would decrease the amount of compensation available to innocent victims through the tort system. Personally, I think this is more a reason for the State to face up to its distributive obligations than it is a reason to reject (P). But I accept that, in an imperfect world, it might be preferable, all things considered, to make do with an incoherent tort system. That doesn't make it any less incoherent.

The second main objection to (P) is that it fails to adequately punish defendants. I claimed that in *Fire 6*, Greta can choose to sue either Big Corp or Little Corp for 100% of her losses, after which the other company's second-order duty disappears. Hence if Greta sues Big Corp, Big Corp would end up paying 100% of Greta's losses and Little Corp would end up paying nothing. This seems unfair. After all, Little Corp is just as

much *to blame* for the damage to Greta's house as Big Corp is. It seems wrong to allow Little Corp to get away with its actions just because Greta chose to sue someone else.[35]

I agree that there is *something* unfair about Little Corp getting away with its negligent behaviour. But this is a *retributive* injustice, not a corrective one.[36] To illustrate, consider a case just like *Fire 6*, except that the fire started by Big Corp provably burnt down Greta's house before the fire started by Little Corp arrived on the scene. Most people would agree that in this case, the damage to Greta's house was caused by Big Corp's negligence alone – Little Corp's negligence was *pre-empted* by Big Corp's negligence, so that Little Corp's actions in fact played no role in bringing the damage about. (WC), (T) and (P) would all therefore imply that Greta has no claim against Little Corp in such a case. From a *retributive* perspective, it might seem unfair that Little Corp has escaped financial punishment and Big Corp has not – it was clearly only a matter of *luck*, after all, that the house had already burnt down by the time Little Corp's fire arrived. But tort law is not a system of retributive justice. The demands of tort law might *coincide* with the demands of retributive justice in some cases, but this is merely a side-effect. There are other systems, such as the criminal justice system, which play the role of delivering the demands of retributive justice already (or at any rate, there ought to be).

I think the same considerations apply to *Fire 6*. It's not part of the remit of tort law to ensure that Little Corp is punished *Fire 6*, no matter how blameworthy it might be. Once Big Corp has compensated Greta for her losses, the demands of corrective justice have been met, and Little Corp's second-order duty of repair evaporates. How Little Corp ought to be punished, if at all, for wrongfully causing Greta's harm is an important question, but it's a question of retributive justice, not corrective justice, and

[35] Thanks to Sara Bernstein for pressing this objection.

[36] One can also see this as a distributive injustice – the costs of repairing Greta's house have been distributed unfairly. See above for my response to this worry.

hence not a question to which tort law need provide an answer. Little Corp may well have committed a *crime*, as well as a tort, for which it may still be prosecuted, even after civil proceedings have come to an end – if so, it is the responsibility of the criminal justice system to ensure that retributive justice is done.

Again, some people might insist that, even if it's fundamentally a system of corrective justice, tort law can still legitimately be used to pursue retributive goals, perhaps in order to pick up some of the slack created by the state's failure to address them. Some jurisdictions, for example, allow claimants to sue for *punitive* damages, over and above ordinary damages, in cases where the defendant's conduct is judged to be particularly insidious.[37] Punitive damages very obviously have nothing to do with meeting the demands of corrective justice. Rather, "an award of punitive damages expresses the community's abhorrence at the defendant's act…[and] commutes our indignation into a kind of civil fine, civil punishment".[38]

But again, the danger with this approach is that we end up with a tort system which cannot be justified on either corrective or retributive grounds. Suppose that Big Corp's negligence in *Fire 1* was particularly insidious. Perhaps Big Corp had known all along about the risks its outdated equipment posed to Greta, but deliberately and cynically ignored them. Greta is awarded compensatory damages to cover her losses and an extra sum in punitive damages. The extra sum is justified on retributive grounds, because the compensatory damages were insufficient to properly punish Big Corp for its particularly blameworthy behaviour. But why should *Greta* receive the punitive

[37] Punitive damages are an established feature of case law in the USA and to a lesser extent Canada. For guidelines, see *BMW of North America v. Gore*, 517 U.S. 519, (1996); *Whiten v. Pilot Insurance* [2002] 1 SCR 595. They are much less common elsewhere, though they are allowed in exceptional cases, for lower sums, in Australia – see *Gray v. Motor Accident Commission* (1998) 196 CLR 1 – and New Zealand, despite the abolition of compensatory tort damages with the creation of the Accident Compensation Corporation – see *Auckland City Council v. Blundell* [1986] 1 NZLR 732.

[38] *Kemezy v. Peters*, 79 F. 3d 33, 35. This judgement was explicit about the fact that one of the functions of punitive damages is to "relieve the pressures on the criminal justice system".

damages? Her losses have already been repaid. There are many victims who can't recover through the tort system at all; why shouldn't the punitive award go to them instead? Some have suggested that "[p]unitive damages…are awarded to the injured party as a reward for his public service in bringing the wrongdoer to account".[39] But this is difficult to sustain; after all, members of the public who assist the police in bringing *criminals* to account tend not to receive million or billion dollar rewards for their efforts, except perhaps in exceptional cases.

Normative coherence in tort law can only be achieved by using it purely to pursue the demands of corrective justice, and this means replacing (T) with (P) across the board. Holding everything else fixed, such a change would probably increase the level of retributive injustice in the world, since some wrongdoers would receive less punishment than they deserve. Personally, I think this is more a reason for the State to face up to its retributive obligations than it is a reason to reject (P). But I accept that, in an imperfect world, it might be preferable, all things considered, to make do with an incoherent tort system. That doesn't make it any less incoherent.

Similar considerations apply to more subtle objections too. (P) implies that degrees of liability for losses are purely a function of degrees of contribution to causings of the harms from which the losses arose. But even those who reject Wright's (1988: 1146) insistence that "[t]rue 'causal apportionment' is conceptually meaningless" tend to think that calculations of degrees of liability "must have regard to the *blameworthiness* of each party"[40] as well as the contribution each made to the causing of the harm.[41] The standard authority in England is *Davies v. Swan Motors*:

---

[39] *Neal v. Newburger Co.* 154 Miss. 691, 700 (1929).

[40] *Stapley v. Gypsum Mines Ltd* [1953] AC 663, 682 (emphasis added).

[41] Similarly, most US states apportion damages in such cases according to 'relative fault', which is typically taken to be a function of factors like "the significance of what the actor was seeking to attain by his conduct…the actor's superior or inferior capacities", and so on (*Uniform Comparative Fault Act*, s 2)

> [Apportioning liability] involves a consideration not only of the causative potency of a particular factor, but also its blameworthiness. The fact of standing on the steps of a dust cart is just as potent a factor in causing damage, whether the person standing there be a servant acting negligently in the course of his employment or a boy in play or a youth doing it for a lark: but the degree of blameworthiness may be very different.[42]

This passage suggests that if D1's actions contributed to the causing of P1's harm to exactly the same degree as D2's actions contributed to the causing of P2's harm, it might nevertheless be right in some cases to hold D1 liable for P1's losses to a greater degree than D2 is held liable for P2's losses, if D1's actions were more blameworthy than D2's. This is inconsistent with (P).

I think we shouldn't incorporate considerations of blameworthiness into calculations of degrees of liability. Considerations of blameworthiness are obviously relevant to the sentence a defendant can expect to receive in a *criminal* case – murder is punished more severely than manslaughter, for example, and the age of the defendant can be a mitigating factor. This is at it should be, since blameworthy actors deserve to be punished more severely than less blameworthy ones. But tort law is not a system of retributive justice. In standard tort cases, where the defendant's wrongdoing caused the claimant's harm, the defendant cannot cite her age as a mitigating factor and expect to have her degree of liability reduced – she would be required to pay 100% of the claimant's losses, whatever the sum, *regardless* of her age, whether the wrong was intentional, reckless or merely negligent, and so on. Punitive awards notwithstanding, the value of the damages awarded in standard tort cases, assuming the claim is successful, is purely a function of the losses incurred, and not at all a function of the blameworthiness of the defendant's actions. Hence I see no reason why considerations of blameworthiness should be incorporated into calculations of degrees of liability in

[42] *Davies v. Swan Motors* [1949] 2 KB 291, 326.

any tort case. To do so, in my view, is again to confuse tort law for a system of retributive justice, which it is not.

There is one final objection to (P) we should consider. Suppose Qadir wants to kill Robert. He knows, however, that he will be held liable to Robert's next of kin for losses arising from Robert's death if he does so. So he persuades Susie to hatch a plan with him – they will each injure Robert non-fatally, but in such a way that Robert's death will be collectively caused by both injuries. (P) seems to imply that this strategy would enable Qadir to successfully reduce his degree of liability for the losses arising from Robert's death. This seems bad. At the very least, you might think, there is something problematic about a tort system that provides an incentive for such behaviour. The problem here seems to be that Qadir and Susie, unlike Big Corp and Little Corp in *Fire 5* and *Fire 6* above, are not *independent* actors. They have a *joint intention* to bring about a particular harm, and their actions are designed to satisfy that joint intention. As a result, it seems inappropriate to hold each person liable only to the degree to which they contributed to the causing of Robert's death.

For those who feel the pull of this intuition, one option is to appeal to the concept of *group agency*. A number of philosophers have recently argued that we should recognise the possibility of group agents, entities which can form intentions, perform actions, and acquire duties that are in some important sense 'over and above' the intentions, actions and duties of their constituent members.[43] One might therefore insist that the conspiracy between Qadir and Susie to kill Robert *together* makes it appropriate to think of them as constituting a single group agent, and to think of their respective actions as constituting a single group action. Applying (P) to this case then

[43] See especially List and Pettit (2011); Pettit (2007, 2009). In arguing for this conclusion, List and Pettit place particular significance on judgement aggregation paradoxes like the so-called 'doctrinal paradox' (Kornhauser and Sager 1993). List and Pettit (2002) generalise the doctrinal paradox to show that there is no complete and consistent way of aggregating individual judgements satisfying certain minimal conditions (subsequent work has succeeded in weakening these conditions even further – see Pauly and van Hees (2006), for example).

yields the simple conclusion that the group agent is fully liable for Robert's death in virtue of performing an action that (individually) caused it.

It has also been argued, albeit more controversially, that there is a kind of liability an individual agent can acquire merely in virtue of being a member of a group which has wrongfully contributed to a causing of harm.[44] If this thesis is correct, the right way of allocating liabilities belonging to a group amongst the group's constituent members will plausibly depend on factors – such as the position of the individual with the power hierarchy of the group, for example – which have nothing to do with the individuals' own degrees of contribution to the causing of the harm. In this way, Qadir may find himself with a degree of liability for the losses arising from Robert's death that is larger than the degree of liability he has merely in virtue of the contribution his own individual action made to the causing of the death. This sketch of a response raises some important questions, of course. Under what conditions does a mere collection of individuals come to constitute a group agent? By what mechanism does an individual 'inherit' liability for wrongs committed by a group of which she is a member? Interesting though they are, these are questions for another time. The important point is that it is possible to account for the moral and legal significance of agents acting with a *joint* intention to bring about a harm in a way that is consistent with (P).

7. Conclusion

In this paper, I have defended a causal theory of partial liability – a defendant should be held liable for a claimant's losses only to the degree to which the defendant's wrongdoing contributed to the causing of the claimant's harm. Though this view has some revisionary consequences for tort law as actually practiced, I've argued that they

[44] See especially, Feinberg (1968); May (1992). This question has historically been tied up with whether individuals can rightly be held responsible for war crimes or atrocities perpetrated by groups of which they are members – see Jaspers (1961) and Levinson (1973) on the Nuremberg trials, for example – but the debate has consequences for tort law too.

can be defended, at least on a conception of tort law as a system of corrective – and not distributive or retributive – justice.

Some people might agree with everything I've said *in theory*, but worry that, in practice, adopting (P) across the board would lead to chaos. "Causation itself is difficult enough; degrees of causation would really be a nightmare." (Chapman 1948: 28). There are understandable Rule of Law concerns – determining degrees of causal contribution is difficult, and different judges and juries might come to entirely different conclusions on the same facts. Potential claimants who know the law might still therefore be ignorant of the likely outcome of legal action.

I have some sympathy for these concerns.[45] There is the question of how things ought to be in a perfect world ('ideal theory'), and then there is the question of how things ought to be, holding fixed the various imperfections of the world in which we live ('non-ideal theory'). This paper has been an exercise in ideal theorizing. I have defended a vision of what tort should aim to be. I invite others more qualified than I to discuss any consequences my arguments may have for the more practical question of how things ought to be, holding fixed the limitations on resources and time and evidence that govern actual tort cases.

References

Barker, K. and Steele, J. (2015). Drifting towards proportionate liability: Ethics and pragmatics. *The Cambridge Law Journal 74*(1), 49-77.

Beale, J. H. (1920). The proximate consequences of an act. *Harvard Law Review 33*(5), 633-658.

[45] It's worth noting, however, that judges *already* have complete discretion over how to 'apportion' damages in cases with multiple wrongdoers, even in joint and several liability systems (albeit after the initial claim succeeds).

Bernstein, S. (forthcoming). Causal proportions and moral responsibility. In D. Shoemaker (ed.), *Oxford studies in agency and responsibility*, Oxford: Oxford University Press.

Bigelow, M. M. (1903). *The law of torts*. Cambridge: Cambridge University Press.

Chapman, S. (1948). Apportionment of liability between torfeasors. *Law Quarterly Review 64*(1), 26-45.

Coleman, J. L. (1992). *Risks and wrongs*. Cambridge: Cambridge University Press.

Coleman, J. L. (2001). *The practice of principle*. Oxford: Oxford University Press.

Dretske, F. (1969). *Seeing and knowing*. London: Routledge and Kegan Paul.

Edgarton, H. W. (1924). Legal cause. *University of Pennsylvania Law Review 72*(3-4), 211-244, 343-375.

Epstein, R. A. (1973). A theory of strict liability. *Journal of Legal Studies 2*(1), 151-204.

Feinberg, J. (1968). Collective responsibility. *Journal of Philosophy 65*(21), 674-688.

Gardner, J. (2014). What is tort law for? Part 2. The place of distributive justice. In J. Oberdiek (ed.), *Philosophical foundations of tort law*, Oxford: Oxford University Press.

Gardner, S. (1996). Contributory negligence, comparative negligence, and stare decisis in North Carolina. *Campbell Law Review 18*(1), 1-73.

Green, L. (1927). *Rationale of proximate cause*. Kansas City, MO: Vernon Law Book Company.

Held, V. (1970). Can a random collection of individuals be morally responsible?. *The Journal of Philosophy 67*(14), 471-481.

Hohfeld, W. N. (1917). Fundamental legal conceptions as applied in judicial reasoning. *The Yale Law Journal 26*(8), 710-770.

Hurd, H. M. and Moore, M. S. (2002). Negligence in the air. *Theoretical Enquiries in Law 3*(2), 333-411.

Jaspers, K. (1961). *The question of German guilt (trans. E. B. Ashton)*. New York: Capricorn.

James, F. and Perry, R. F. (1951). Legal cause. *The Yale Law Journal 60*(5), 761-811.

Kornhauser, L. A. and Sager, L. G. (1993). The one and the many: Adjudication in the collegial courts. *California Law Review 81*(1), 1-59.

Levinson, S. (1973). Responsibility for crimes of war. *Philosophy and Public Affairs 2*(3), 244-273.

List, C. and Pettit, P. (2002). Aggregating sets of judgments: An impossibility result. *Economics and Philosophy 18*(1), 89-110.

List, C. and Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford: Oxford University Press.

Malone, W. (1956). Ruminations on cause-in-fact. *Stanford Law Review 9*(1), 60-99.

May, L. (1992). *Sharing responsibility*. Chicago: The University of Chicago Press.

McNichols, W. J. (1979). Judicial elimination of joint and several liability because of comparative negligence – A puzzling choice. *Oklahoma Law Review 32*(1), 1-62.

Moore, M. S. (2009). *Causation and responsibility*. Oxford: Oxford University Press.

Pauly, M. and van Hees, M. (2006). Logical constraints on judgement aggregation. *Journal of Philosophical Logic 35*(6), 569-585.

Pettit, P. (2007). Responsibility incorporated. *Ethics 117*(2), 171-201.

Pettit, P. (2009). The reality of group agents. In C. Mantzavinos (ed.), *Philosophy of the social sciences: Philosophical theory and scientific practice*. Cambridge: Cambridge University Press.

Pollock, F. (1920). *The law of torts (11th edition)*. London: Stevens and Sons.

Pollock, F. (1922). Liability for consequences. *Law Quarterly Review 38*(2), 165-167.

Posner, R. and Landes, W. M. (1980). The positive economic theory of tort law. *Georgia Law Review 15*(4), 851-924.

Ripstein, A. (1998). *Equality, responsibility and the law*. New York: Cambridge University Press.

Rogers, W. V. H. (ed.) (2004). *Unification of tort law: Multiple tortfeasors (Vol. 9).* Kluwer Law International.

Sartorio, C. (forthcoming). A new form of moral luck? In A. Buckareff, C. Moya and S. Rosell (eds.), *Agency, freedom, and moral responsibility,* London: Palgrave Macmillan.

Schroeder, C. H. (1990). Corrective justice and liability for increasing risks. *UCLA Law Review 37*(3), 439-478.

Seavey, W. A. (1939). Mr. Justice Cardozo and the law of torts. *Harvard Law Review 52*(3), 372-404.

Smith, J. (1911). Legal cause in actions of tort. *Harvard Law Review 25*(2), 103-128.

Stapleton, J. (2001). Legal cause: Cause-in-fact and the scope of liability for consequences. *Vanderbilt Law Review 54*(3), 941-1024.

Steel, S. (2012). Causation in English tort law: Still wrong after all these years. *University of Queensland Law Journal 31*(2), 243-264.

Steel, S. (2015). *Proof of causation in tort law*. Cambridge: Cambridge University Press.

Steel, S. and Ibbetson, D. (2011). More grief on uncertain causation in tort. *The Cambridge Law Journal 70*(2), 451-468.

Strudler, A. (1992). Mass torts and moral principles. *Law and Philosophy 11*(4), 297-330.

Summers, R. S. (1982). *Instrumentalism and American legal theory*. Ithica, NY: Cornell University Press.

Weinrib, E. J. (1995). *The idea of private law*. Cambridge, MA: Harvard University Press.

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.

Wright, R. W. (1985). Causation in tort law. *California Law Review 73*(6), 1735-1828.

Wright, R. W. (1988). Allocating liability among multiple responsible causes: A principled defense of joint and several liability for actual harm and risk exposure. *University of California at Davis Law Review 21*(4), 1141-1211.