Criminal Law and Philosophy (2024) 18:151–166
https://doi.org/10.1007/s11572-023-09694-4

ORIGINAL PAPER

Moore on Degrees of Responsibility

Alex Kaiserman[1]

Accepted: 4 July 2023 / Published online: 17 July 2023

Abstract
In his latest book *Mechanical Choices*, Michael Moore provides an explication and defence of the idea that responsibility comes in degrees. His account takes as its point of departure the view that free action and free will consist in the holding of certain counterfactuals. In this paper, I argue that Moore's view faces several familiar counterexamples, all of which serve to motivate Harry Frankfurt's classic insight that whether and to what extent one is responsible for one's action has more to do with what *actually caused* that action than with what one could or couldn't have done instead. I then go on to sketch an alternative approach to degrees of responsibility that takes seriously this insight. I'll argue that Moore ought to be sympathetic to this approach, inasmuch as it combines two familiar Moorean ideas: the idea that causal contribution comes in degrees, and the idea that acting freely is compatible with, and indeed entails, the fact that one's action was caused by prior states of affairs.

Keywords Responsibility · Free will · Diminished responsibility · Counterfactuals · Causation

1 Introduction

Having presented his compatibilist theory of responsibility in Part IV of *Mechanical Choices*, Moore pauses to note a potentially "disquieting implication" of his analysis: namely, that it is "incapable of unpacking a *binary* distinction" between responsible and non-responsible conduct, replacing it instead with "a scalar distinction, one whereby [responsibility] is a more-or-less affair, a matter admitting of degrees"

✉ Alex Kaiserman
alexander.kaiserman@philosophy.ox.ac.uk

[1] Fairfax Fellow and Tutor in Philosophy, Balliol College, University of Oxford, Broad Street, Oxford OX1 3BJ, UK

(p.355).[1] On reflection, Moore embraces this conclusion as "a virtue, not a vice, of the analysis" (p.356). "As a legal matter the criminal law may impose a binary categorization on this by-degree continuum… But the law here does no more than what it does in many places, which is attach a bivalent remedy on what we all know is in nature a matter of continuous variation" (ibid.).

Moore is highlighting an important but neglected feature of our responsibility practices: often we feel intuitively that a person should be partially, but not fully, excused for their conduct, on account of them being partially, but not fully, responsible for it. Consider the facts of *R v. Campbell*,[2] in which a hitchhiker was killed by a man who (it was later revealed on appeal) had frontal lobe damage caused by epileptic seizures that substantially impaired his ability to control his emotions, process information, and appreciate the significance of his actions. In this case and many like it, the appropriate response seems to lie somewhere between unmitigated blame and full excuse, with the exact point on the scale depending, in part, on the extent of the impairment. Indeed, most of the conditions the law recognizes as excusing – insanity, involuntary intoxication, immaturity, and so on – can seemingly obtain to a greater or lesser extent.

This fact has long been recognized by lawyers, if not always by the law itself. As early as the 17th Century, Matthew Hale acknowledged that "[t]here is a partial insanity of mind…some persons that have a competent use of reason in respect of some subjects, are yet under a particular dementia in respect of some particular discourses, subjects or applications; *or else it is partial in respect of degrees*" (Hale 1736: ch.iv; my emphasis). Despite this though, Hale argued that "partial insanity seems not to excuse [a person] in the committing of any offence" (ibid.). Hale's contemporary in Scotland, George Mackenzie, took a different view: "Since the law grants a total impunity to such as are absolutely furious, it should by the rule of proportions lessen and moderate the punishments of such, as though they are not absolutely mad yet are Hypochondrick and Melancholy to such a degree, that it clouds their reason" (Mackenzie 1678: pt.1 tit.1 sc.8). Walker (1968: ch.9) identifies this as the moment the law in Scotland departed from that in England. Although the defence of insanity was technically all-or-nothing, there developed a practice, in cases where the standard penalty was execution, of allowing (or in some cases even advising) the jury to return a verdict of guilty with a recommendation for royal mercy. This culminated in Lord Deal's celebrated innovation in *H M Advocate v. Dingwall*, in which he invited the jury to reduce the defendant's conviction from murder to culpable homicide (for which there was more flexibility in sentencing), in recognition of the fact that his "state of mind [was] an extenuating circumstance, although not such as to warrant an acquittal on grounds of insanity";[3] thereby effectively transferring to the jury a power that had previously belonged only to the royal prerogative. A defence of diminished responsibility is now widely recognized in common law jurisdictions,[4]

[1] Page references are to *Mechanical Choices* unless indicated otherwise.

[2] [1987] 84 Cr App R 255.

[3] (1867), 5 Irvine 466.

[4] The practice of empowering juries to recommend a mitigated sentence in cases involving 'partially insane' defendants had obvious attractions to 19th -Century English lawyers, one of whom even sug-

and its development is an instructive case study in what happens when the law tries to impose binary distinctions on what ordinary judgement recognizes is an underlying moral continuum.

So I agree with Moore that responsibility comes in degrees. But as we'll see, I disagree with him on how degrees of responsibility should be understood. In this paper I will mostly focus on these points of departure. I start in Section 2 with a careful exposition of what I take Moore's view to be, distinguishing it from other views with which I think it is sometimes conflated in *Mechanical Choices*. Section 3 raises a general challenge for Moore's view, which draws on Frankfurt's (1969) classic insight that whether and to what extent one is responsible for one's action has more to do with what *actually caused* that action than with what one could or couldn't have done instead. Section 4 sketches an alternative approach to degrees of responsibility that takes this insight as its starting point. I'll argue that Moore ought to be sympathetic to this approach, inasmuch as it combines two familiar Moorean ideas: the idea that causal contribution comes in degrees, and the idea that acting freely is compatible with, and indeed entails, the fact that one's action was caused by prior states of affairs.

2 Moore's Counterfactual Account of Freedom

To understand Moore's account of degrees of responsibility, we first need to understand his theory of action. Moore defends a traditional 'tripartite' theory of rational action, whereby a desire that *p* and a belief that φ-ing would make it the case that *p* combine to produce an intention to φ, which itself causes the agent to φ. When an agent satisfies this schema in her reasoning and action, "there is not even a hint of excuse…hers is a paradigm of rational action for which she is fully responsible" (p.328). Of course in reality things are usually more complicated than this, because we usually have multiple desires that bear on the question of what to do in some situation, which may conflict in what course of action they recommend. For example, I may want to eat the cake, but also want to stick to my diet. In such cases, Moore thinks, there will usually be a 'strongest' desire, which he defines operationally as the desire that would win out in causing the agent's choice were the conflict known to the agent and resolved through the formation of a conflict-resolving intention.[5] When an agent acts on their strongest desire, so-understood, "there is again no hint of excuse" (p.329).

Thus the question of whether to hold someone responsible for their conduct only truly arises, for Moore, in cases where an agent fails to do what they believe will bring about the situation they most desire. This could happen for one of two reasons:

gested that "the law ought…to allow the jury to return any one of three verdicts: Guilty; Guilty, but his power of self-control was diminished by insanity; Not Guilty on the ground of insanity" (Stephen 1883: 175). But parliamentary gridlock prevented any such reforms, until a defence of diminished responsibility was finally introduced in the 1957 Homicide Act. Note that the defence is still only available in murder cases, however – for an argument that there should be a generic partial excuse of diminished responsibility, regardless of the crime, see Morse (2003).

[5] Moore appears to use 'choice' and 'formation of an intention' interchangeably.

either the agent fails to form an intention to act on their strongest desire (which, given how Moore defines 'strongest desire', means that they must either have failed to recognize the conflict in their desires or failed to form an intention to resolve it), or they form an intention to act on their strongest desire but fail to act on it. Yet even in such cases, Moore thinks, the agent would still be responsible for their action so long as they *could* have acted and chosen to act on their strongest desire. We only excuse those who can't do better, not those who merely won't.

The question that now arises is what is meant by 'can' in the claims above. Here Moore follows his namesake (Moore 1912) in defending a *conditional* analysis of the agential 'can', albeit with a contextualist twist – a particular utterance of the form 'S could have φ-ed' expresses the proposition that S *would* have φ-ed had conditions C obtained, with C being determined by the context. For example, if a coach says to his track star, 'You could have won that race', she might mean that he would have won it had he trained harder, or that he would have won had he gotten more sleep the night before, or any number of other things, depending on the context.

This approach raises a further question, though – if 'can' is context-sensitive, what is the particular sense of 'can' that is relevant to freedom and responsibility? Here two 'extreme' answers present themselves. At one extreme is the 'ultraliberal', according to whom one is responsible for one's action only if one could have acted otherwise *given* the particular causal histories of one's actions and choices. On this view, what is revealed by the (alleged) fact that our choices and actions are causally determined by past neurophysical and environmental factors over which we have no control is that we couldn't, in the relevant sense, have chosen or acted otherwise than we actually did, and hence that none of our actions or choices are free. At the other extreme is the 'ultraconservative', according to whom one is responsible for one's action so long as there are *some* conditions, perhaps very different from those that actually obtained, under which one would have chosen/acted otherwise than one actually did. Thus if one would have refrained from committing a crime had there been 'a policeman at one's elbow', one is responsible for committing it, on this view.

Moore's aim in this chapter is to trace a middle way between these two extremes, to find a version of compatibilism that doesn't 'overshoot' by making us responsible for too much. He does this by stipulating interpretations of the 'free action' condition (that one is responsible for one's action only if one could have acted otherwise) and the 'free will' condition (that one is responsible for one's action only if one could have chosen to act otherwise) that he believes "tame compatibilism, by keeping its analysis of incapacity within morally plausible dimensions" (p.355). Starting with the free will condition, Moore's view is that S could have chosen to act otherwise, in the relevant sense, if and only if S would have chosen to act otherwise *had they wanted to do so badly enough*. If I would have chosen to eat the slice of cake despite wanting very much to stick to my diet – for example, because my desire for cake is of a kind that refuses to 'integrate' with my other desires in a way that would give rise to a conflict-resolving intention – then I couldn't, in the relevant sense, have chosen not to eat the cake, and I am not responsible for my action.

Moore's proposed interpretation of the free action condition is a bit more subtle. His first pass is this: S could have acted otherwise, in the relevant sense, if and only if S would have acted otherwise *if they had chosen to do so*. But Moore recognizes

that this won't do, since there are cases where an agent fails to do what they choose to do, thereby trivially violating the condition above, but intuitively are still responsible for their action. Typically this happens because the agent's intention is 'non-sticky' – subject to frequent re-evaluation of what the agent most desires. For example, I might at one time desire most to stick to my diet, and correspondingly choose not to eat the cake; but shortly afterwards desire most to eat the cake, form the opposite intention, and eat the cake. At all times I intend in line with what I most desire, but the relative strength of my desires fluctuates. Sometimes this (diachronic) weakness of will is simply a moral defect in an agent's character, and in these cases they shouldn't be excused. But sometimes – for example if the agent is involuntarily intoxicated, or very young, or mentally ill – the non-stickiness is not the agent's fault, and in these cases they are not responsible for their actions. To deal with this, Moore suggests that an agent could have acted otherwise, in the relevant sense, if and only if they would have acted otherwise had they formed an intention to act otherwise with "the minimal stickiness required by morality for a person of [their] type" (p.354).

It's at this point that the "disquieting implication" of Moore's analysis looms into view. As Moore notes, there is an element of degree vagueness in the phrases 'wanted to do so badly enough' and 'minimal stickiness required by morality'. But this, we're told, is really "a virtue, not a vice, of the analysis" (p.356). Fundamentally, "capacity/incapacity is a matter of degree… We can only say that some have *more* capacity to have done otherwise on a given occasion, and some have less" (ibid.). More precisely, Moore's view appears to be this: the degree to which S could have acted otherwise is proportional to the minimum amount of 'stickiness' such that if S had had an intention to act otherwise with that level of stickiness, she would have done so; and the degree to which S could have chosen otherwise is proportional to the minimum strength of desire such that if S had desired to that extent to choose otherwise, she would have done so. S's degree of responsibility for her action, then, is presumably some function of both these quantities. For practical purposes we might say, on some occasion, that "[i]f the strength of the controlling desire or the stickiness of the executing intention had been 'a lot greater' and yet the accused still would have done what she did, then the level of excuse is reached" (p.357). But this can amount to no more than the imposition of a vague threshold on what is in reality an underlying moral continuum; "[a]nd it takes some seemingly arbitrary stipulation to say where on that matter of continuous variation 'can't' begins and 'won't' ends" (p.356).

In the next section I will critically assess Moore's account of degrees of responsibility, as I have reconstructed it here. But before we move on I want to flag one distracting feature of how Moore introduces his view which I have deliberately ignored up until now. As Moore notes, counterfactuals are typically analysed in terms of possible worlds; and at various points, Moore appears to suggest that the scalarity of capacity is a *consequence* of the possible-worlds analysis of counterfactuals.[6] On its face, this claim is puzzling. According to the standard Lewis-Stalnaker semantics (where '>' denotes the counterfactual conditional), $p > q$ is true if and only if q is true

[6] "[T]he possible worlds analysis is incapable of unpacking a *binary* distinction between what one can't do versus what one doesn't do" (p.355); "If the possible worlds analysis of the counterfactuals involved with capacity is correct…capacity/incapacity is a matter of degree" (p.356); etc.

in all the closest worlds in which p is true.[7] How these truth-conditions are supposed to give rise to scalarity is not immediately clear. Indeed, the truth-conditions of $p > q$ on the standard analysis are *absolute*, in two senses: only the *closest* p-worlds matter, and q must be true in *all* of them.

Later though we get a better idea of what Moore has in mind: "The closer the possible worlds in which someone would have chosen or done otherwise, the more ability he had to choose or do otherwise in this, the actual world; the more remote the possible worlds in which someone still would not have chosen or acted other than they did, the less ability he had to choose or do otherwise in this, the actual world" (p.356).[8] Curiously, Moore appears to think that this account of degrees of capacity in terms of closeness of worlds is equivalent to the account in terms of strength of desire and stickiness of intention described above. But it isn't – there is no reason in general to think that the worlds in which an agent's desire and intention to act otherwise are stronger or stickier must thereby be further away from actuality.[9] To illustrate, suppose that Alan gives in to temptation and eats a slice of cake. Now suppose that, unbeknownst to Alan, Becky has poisoned Alan's cake; she has recently come to regret this, and came very close to confessing to Alan, but at the last moment decided against it. Adding Becky to the story makes the worlds where Alan acts otherwise closer to the actual world – Becky *could easily* have told Alan the cake was poisoned, and if she had, Alan wouldn't have eaten it. But plainly it should make no difference to Alan's degree of responsibility for his action. How able Alan was in *this* world to stick to his diet shouldn't be sensitive to how close someone came to telling him the cake was poisoned. By contrast, the view I attributed to Moore above gets the right result here, because making those worlds in which Alan has a strong desire to act otherwise closer to actuality doesn't change the strength of desire that would have been required for him to act differently.[10]

[7] I'm assuming here that there is always a closest world in which p is true; this assumption is controversial, but the details won't matter for our purposes.

[8] As stated this can't be right, since it makes degrees of capacity sensitive to the modal robustness of our choices and desires – if I simply really like cake, so that the worlds where I choose not to eat cake are far away from actuality, it shouldn't automatically follow that my ability not to eat the slice of cake on this occasion is diminished, or that I bear only a small degree of responsibility for my action. A better view (following a suggestion I make (Kaiserman 2021) on behalf of a similar view due to Coates and Swenson (2013)) would define one's degree of freedom of action as the distance from actuality of the closest worlds in which the agent chooses to act otherwise and does so, *as a fraction of* the distance from actuality of the closest worlds in which the agent chooses to act otherwise. This function returns the value 1 if the agent would have acted otherwise had she chosen to, and tends to 0 the further away from the closest world in which the agent chooses otherwise one has to go to find one in which she acts in line with her choice. (*Mutatis mutandis* for the free will condition.)

[9] Indeed Moore acknowledges this very point earlier in the chapter: "In light of the reverberations that any change in strength of desire will have on other states and the laws that connect them, closeness of possible worlds in which we judge whether X would have chosen differently do not just depend on the differential strength of the controlling desire and on the mechanisms through which even very strong controlling desires may fail to determine choice. If a small change in the strength of controlling desire requires large changes in the causes of such strength, or in scientific laws, or both, such small change may not betoken a close possible world" (p.352).

[10] There are, in general, two ways of adapting counterfactual definitions like Moore's to accommodate scalarity in terms of possible worlds. Suppose $p > q$ is false; still, there are two different senses in which it might be 'nearly' true: (i) the closest worlds where p and q are both true are *only a bit further away* than

I conclude that, despite Moore's apparent remarks to the contrary, his account of degrees of responsibility in fact has little to do with possible-worlds semantics of counterfactuals. What's important, on Moore's view, are the stickiness of one's intention to act otherwise and the strength of one's desire to choose to act otherwise – one's degree of responsibility is proportional to the minimum points on those scales at which the relevant counterfactuals become true (regardless of how counterfactuals generally should be analysed).

3 Masks, Finks and Frankfurt Cases

In this section, I want to raise a general challenge to Moore's counterfactual approach to degrees of free will and free action. The challenge is not new – it concerns cases variously known as 'Frankfurt-style cases', or 'pre-emption cases'; cases involving back-up mechanisms that play no causal role in bringing about the agent's action or choice but that would have prevented them acting or choosing otherwise under certain specified conditions. It will ultimately be argued that Moore's account fails "in the same way that counterfactual analyses usually fail, by ignoring side-effects of the conditional's antecedent on the truth-value of the analysandum" (Williamson 2000: 209).

Let's begin with a classic Frankfurt-style case. Suppose that Alan chooses to eat a slice of cake, and does so, despite desiring most strongly to stick to his diet. Let's stipulate that Alan has a high degree of responsibility for his action – one can fill in the details however one likes. Now suppose that an evil neuroscientist is monitoring Alan's brain activity. Should she detect signs of Alan having a strong desire to choose not to eat cake, she will intervene (by stimulating the relevant neurons) to ensure that he chooses to eat the cake regardless; as it happens, though, Alan does not have such a desire, and so the neuroscientist does nothing. Intuitively, adding the neuroscientist to the story makes no difference at all to Alan's degree of responsibility for his action, given that her presence is irrelevant to why Alan actually chose and acted as he did. Yet it dramatically changes the counterfactual structure of the case: whereas without the neuroscientist Alan may well have chosen to act otherwise had his desire to do so

the closest worlds where p is true; or (ii) q is true in *nearly all* the closest worlds where p is true. These suggest two alternative ways of thinking about S's degree of free will, in terms of either the distance from actuality of the closest worlds where S wants to choose otherwise and does so (as a fraction of the distance from actuality of the closest worlds where S wants to choose otherwise), or in terms of the fraction of those worlds in which S wants to choose otherwise where she does so. (Similar choice points exist when thinking about the scalarity of dispositions (Vetter 2015), reasons-sensitivity (Kaiserman 2021) and causal contribution (Kaiserman 2018).) Though Moore mostly takes the former approach, there are hints of a commitment to the latter approach in *Mechanical Choices* as well (see for example Moore's comments in Chap. 11 about the "strength of necessitation" between one's choices and the neurophysical changes which typically precede them, which, notwithstanding Moore's explicit reliance "on an intuitive equation of strength of necessitation with closeness of possible worlds", seems more suggestive of a view on which one's degree of responsibility depends on the *number* of nearby possibilities in which one's choice is accompanied – whether afterwards or beforehand – by the relevant action, rather than their closeness to actuality).

been just a little stronger, with the neuroscientist he wouldn't have done so regardless of how strong his desire had been. Moore's view thus gets the wrong results.[11]

There are also cases with the opposite structure. Suppose again that Alan eats a slice of cake, but this time let's stipulate that he has a low degree of responsibility for his action. Again, one can fill in the details however one likes – perhaps Alan's action is the result of a serious compulsive disorder, for example. Now suppose that a benevolent (/paternalistic) neuroscientist is monitoring Alan's brain activity and, should she detect signs of Alan having a strong desire to choose not to eat cake, will intervene (for example, by rewiring his brain to cure him of his compulsive disorder) to ensure that he chooses not to eat it. As it happens, though, Alan has no such desire and the neuroscientist does nothing. Intuitively, adding the neuroscientist to the story again makes no difference to Alan's degree of responsibility for his action, despite drastically changing the minimum strength of desire such that if Alan had had a desire of that strength to choose not to eat the cake, he would have done so.

Moore does acknowledge problems of this kind. One solution he suggests is to amend his counterfactual analysis of ability along the lines suggested by Lewis (1997) and Vihvelin (2004). On the resulting view, S could have chosen at *t* to act otherwise (in the relevant sense) if and only if there is a time *t`* and a set of intrinsic properties P1,…, Pn which S has at *t*, such that if S had wanted at *t* to choose otherwise badly enough *and* retained P1,…, Pn until *t`*, their desire and P1,…, Pn would jointly have been an S-complete cause of S's acting otherwise (where a cause of something is 'S-complete' if it is sufficient for it "in so far as havings of properties intrinsic to [S] are concerned, though perhaps omitting some events extrinsic to [S]" (Lewis 1997: 156)). "An unlovely mouthful!", Lewis (1997: 157) admits. But it gets the right results in the cases above. For example in the first case, although Alan wouldn't have chosen not to eat the cake had he badly wanted to do so, nevertheless there are certain intrinsic features of his brain – namely those that the neuroscientist would have changed had she intervened – such that had he badly wanted to choose otherwise *while retaining* those features, his desire and those features would jointly have been an Alan-complete cause of his choosing otherwise.

The problem with this strategy is that masks and finks can themselves be intrinsic to an object.[12] To illustrate, consider the following case.[13] Suppose Alan has a very serious compulsive disorder – serious enough that, if it had caused his decision to eat the slice of cake, we'd be minded to excuse him, at least partially. But now let's stipu-

[11] Notice that the inclusion of the neuroscientist also makes a drastic difference to the distance from actuality of the closest worlds in which Alan refrains from eating cake, so this case would also work as a counterexample to the alternative account in terms of closeness which (as we saw in the previous section) Moore occasionally endorses.

[12] The significance of intrinsic finks and masks for counterfactual accounts of dispositions is discussed by Clarke (2008), Everett (2009), Tugby (2016) and Choi (2012), among others. A plausible move in that debate is to simply deny that objects possessing intrinsic masks or finks actually have the corresponding dispositions allegedly masked or finked in the first place; but as Cohen and Handfield (2007) note, the same move is much less plausible when it comes to counterfactual accounts of ability. This lends further support to the view that abilities are not in fact any kind of disposition, notwithstanding the similarities between them (Vetter 2019).

[13] I discuss this case in Kaiserman (2022), but the same strategy is also employed in Cohen and Handfield (2007).

late that Alan's disorder in fact played no causal role whatsoever in bringing about Alan's action – instead Alan was motivated to eat cake by his own reasons, in exactly the way a fully responsible person would have been. Had Alan wanted very badly not to eat the cake, though, his latent compulsive disorder would have been triggered and would have caused him to choose to eat the cake anyway. I think Alan is fully responsible for his action in this case, notwithstanding the fact that he wouldn't have chosen to act otherwise no matter how badly he'd wanted to. And this time Lewis's fix is no help: there are no intrinsic properties of Alan such that had he wanted badly to choose otherwise and retained those properties, his desire and those properties would have been an Alan-complete cause of his choosing otherwise.

The strategy here appears quite general. As I argue in Kaiserman (2022), Frankfurt never actually provided a specific counterexample to the 'principle of alternate possibilities' in his seminal paper. Instead he provided a *recipe* for generating such counterexamples, which is guaranteed to work *however* we choose to interpret the phrase 'could have done otherwise' (which Frankfurt (1969: 834) accepts is highly context-sensitive). First we invite our opponent to describe a case in which someone (call them A) isn't responsible (or is responsible to a low degree) for their action because they couldn't, in whatever sense our opponent deems to be relevant, have acted (or chosen to act) otherwise. Next we ask them to describe a case in which someone (call them B) is responsible (to a high degree) for a similar action. And finally, we simply amend the first case by stipulating that whatever caused A's action is *pre-empted* by a mechanism of exactly the same kind as that which caused B's action in the second case. The result is a case in which A seems perfectly responsible for their action, as responsible as B is in the second case, and yet couldn't have acted otherwise in whatever sense our opponent has seen fit to describe.[14]

4 The Contribution Account

Here's what I take to be the moral of the previous section: how responsible an agent is for their action seems to have much more to do with what *actually caused* their action than it does with what they could or would have done had things been different. The reason why adding inert back-up mechanisms like the neuroscientist makes no difference to Alan's degree of freedom and responsibility in the cases above is precisely that it makes no difference to the causal history of his actions and choices. This suggests a different approach to understanding free will and responsibility, one on which facts about whether and to what extent an agent acted freely are *fully grounded* in the action's causal history.[15] In this section I will describe a view of this kind, which I call the *contribution account* of degrees of responsibility.

[14] The only way of blocking this strategy, I suggest, is by building causal facts themselves into the stipulated interpretation of 'could have done otherwise'. But while this may succeed in reconciling principles like the PAP with our intuitions in Frankfurt-style cases, it is not a dialectically effective strategy, since it succeeds only by helping itself to the very facts its detractors argue are those that ground facts about responsibility in the first place.

[15] For an explanation and defence of this view, see Sartorio (2016). As Sartorio notes, the view is compatible with causal facts themselves being grounded in modal facts.

One can think of the contribution account as combining two elements. The first is a theory of degrees of causal contribution which I have defended elsewhere (Kaiserman 2016, 2017a). On this view, while causation itself is not a scalar relation – X can't cause Y 'a lot' or 'a little' – it is nevertheless a relation to which multiple events can contribute to differing extents – X can contribute a lot, together with other events, to the causing of Y. (In this way *causing* is like *authoring* – one cannot *author* something 'a lot', but one can *contribute* a lot, along with others, to the authoring of a book, for example.) Although this basic insight is compatible with many different accounts of causation, it fits especially neatly with a view according to which causes are *minimally jointly sufficient in the circumstances* for their effects: $X_1,\ldots, X_n$ jointly caused Y if and only if they were jointly sufficient in the circumstances for Y, and no proper sub-plurality of $X_1,\ldots, X_n$ were jointly sufficient in the circumstances for Y.[16] To be *a cause* of Y is to be one of a plurality which collectively caused it (just as to be *an author* of a book is to be one of a plurality of people who collectively authored it). This view explains why causation itself is all-or-nothing – a plurality of events cannot be more or less sufficient for another event – but also why, when multiple events jointly cause an effect, each may *contribute* more or less to the causing of it – they do so by contributing more or less to making the plurality as a whole jointly sufficient for the effect.[17]

The second element of the contribution account is the idea that being responsible for one's action or choice is a matter of it having the *right sorts of causes.* More precisely, some causes of our actions and choices are what I will call *responsibility-grounding* – they *make* the actions and choices they cause free (regardless of what else caused the action/choice, or how the responsibility-grounding causes themselves came about).[18] We can combine this idea with the machinery above as follows. One is *fully* responsible for one's action if it was fully caused by responsibility-grounding factors; one is *partially* responsible for one's action if it was partially caused by responsibility-grounding factors (i.e. if responsibility-grounding factors contributed, perhaps together with other factors, to bringing it about); and one is *not at all* responsible for one's action otherwise. Either way, an agent's degree of responsibility for their action is equal to the maximum[19] degree of contribution responsibility-grounding factors collectively made to a causing of it.

The basic structure of the contribution account can be paired with many different views about what the responsibility-grounding causes are. For example, one might

[16] Such a view faces two challenges: (i) explaining what 'sufficient' means in a way that doesn't make effects sufficient for their causes, or effects of a common cause sufficient for each other; and (ii) explaining what the 'circumstances' mentioned in the analysis are, and if they vary with context, explaining *which* context is the one that matters for attributions of freedom and responsibility. I explore some of these challenges in other work (see especially Kaiserman 2017b), but will largely set them aside here.

[17] Kaiserman (2016) attempts to cash out this idea in probabilistic terms; the details won't be necessary for our purposes, however.

[18] Notice therefore that there is a kind of asymmetry on this view – being responsible for one's action is a matter of it having the right kinds of causes, not a matter of avoiding the *wrong* kinds of causes. See also the discussion of 'selective libertarianism', below.

[19] This qualification is needed because responsibility-grounding factors may contribute to multiple causings of the same action to different extents; see Kaiserman (2016).

take them to be aspects of the agent's *quality of will* (e.g. Björnsson 2017), or aspects of the agent's *deep self* (e.g. Sripada 2016). My own view (Kaiserman 2021), however, draws on Sartorio's (2016) causal account of reasons-sensitivity. Suppose (yet again) that Alan eats a slice of cake. On the one hand, a natural explanation of why adding a causally inert neuroscientist to the case makes no difference to Alan's degree of responsibility for his action is that it makes no difference to the *reasons which motivated him to do it*. Yet on the other hand, two people can seemingly differ as regards how responsible they are for their actions, without differing in the reasons for which they did it. Suppose Alan and Betty both eat a slice of cake for the reason that doing so would satisfy a desire they have for cake; but whereas Alan wouldn't have eaten the cake had there been *very* strong reasons not to do so (e.g. the cake is poisoned), Betty still would have, owing to a pathological inability to integrate her food-related reasons with her other reasons in any coherent kind of way. We want to say that Betty is less responsible for her action than Alan; but this is apparently hard to square with the idea that one's degree of responsibility supervenes on the causal history of one's action.

Sartorio's insight was to see that the causal histories of Alan's and Betty's actions are not in fact the same – the causal history of Alan's action contains various *absences*, which are not part of the causal history of Betty's. For example, Alan's action was partly caused by the absence of anyone poisoning the cake, the absence of anyone threatening to harm him if he eats the cake, and so on, whereas none of these things are causes of Betty's action. This suggests a view on which one is responsible for one's action to the extent to which it was caused by reasons to act that way *and absences of reasons not to act that way*.[20]

Here are two examples to illustrate this view, and how it differs from Moore's. First, consider a drug addict who, at one time, resolves never to take drugs again, but shortly afterwards gives in to temptation. According to Moore, you'll recall, the addict's action is free only if they would have acted otherwise had they had an intention to do so with the minimal 'stickiness' required by morality for a person of their type. But there is something odd about this view, on reflection. After all, how sticky morality requires such an agent to be in their intention not to take drugs will depend on the morality of taking drugs; and on the face of it, whether someone is responsible for taking drugs seems like a question we ought to be able to answer independently of whether taking drugs is morally wrong. Indeed, Moore himself expresses some sympathy, in this chapter and elsewhere, for the view that taking drugs in many cases is morally permissible (see Moore 1998). But if that's right, it's not clear why there would be any moral requirement to stick to one's earlier choice not to take drugs.

[20] There are many different views on what reasons are. My own view is that reasons are *facts* which *tell in favour* of some course of action. But there are different ways in which *p* may tell in favour of S φ-ing. It may do so in a *value-relative* sense – that is, relative to what S herself values – or a *value-independent sense* – that is, relative to what is in fact valuable. And it may do so *objectively* or *subjectively* (where, as I am using the terminology, *p* is a subjective reason to φ if it amounts to evidence of the existence of an objective reason to φ). Probably the most plausible version of the reasons-sensitivity view equates responsible action with action caused by *subjective*, *value-relative* reasons, since this version is compatible with the possibility of one being fully responsible for one's action despite being radically mistaken about either the moral facts or facts about the circumstances. But see Kaiserman (2021) for more on these issues.

Frequently changing one's mind between two morally permissible options might be *irrational*, but it is not immoral. Moreover, even if taking drugs is morally wrong, the question of whether the person going against their earlier choice represents a *moral* failure surely depends on whether the person is responsible, in the first place, for doing so. It seems to get things the wrong way around to say that they are responsible for giving in to temptation *because* their lack of willpower on this occasion is a moral defect.

On my view, how responsible the addict is for taking drugs is independent of the moral status of the action; instead it is a function of how much reasons to take drugs, and absences of reasons not to, contributed to bringing the action about. Perhaps some of those we call addicts take drugs wholly on the basis of reasons to do so, for example that it would alleviate withdrawal symptoms, together with the absence of sufficiently strong reasons not to do so. Such people are, on my view, fully responsible for their actions, whether or not they could have acted otherwise, and regardless of how the reasons for which they acted came about. Moreover, they are fully responsible for their actions regardless of whether the reasons for which they acted were outweighed on this occasion by countervailing reasons not to take drugs, which they ignored – one can be *fully reasons-sensitive* in acting, on my view, without the action being fully *rational*.[21] For other people, taking drugs might not even qualify as an intentional action in the first place – cravings and/or environmental cues cause their actions *directly*, bypassing their reasons entirely. But most addicts will fall somewhere between these two extremes – their actions were partially caused by non-reason-conferring factors, but partially also by absences of strong enough reasons not to take drugs (this is the truth in the claim that many addicts 'maintain some degree of control' over their actions). How responsible such agents are will depend on the relative contributions of each of these factors.

Moore might reply that if an addict takes drugs because of a craving for drugs, having recognized the conflict with their other desires and intentionally resolved that conflict in favour of taking drugs, they are fully responsible for their action, since by definition they act in line with their strongest desire (p.533). But this strikes me as the wrong result, at least in those cases where the craving fails to line up in any sensible way with what the agent most values. As Holton and Berridge (2013) note, a hallmark of addiction is the striking disconnect many addicts display between what they *want* and what they *like*; they desire drugs – not as a means to something else, but intrinsically, for their own sake – despite knowing full well that taking them will provide neither pleasure nor satisfaction. This fact is important on my view, since what matters is whether the action was appropriately caused by *reasons* (and their absences), not desires. Most of the time, admittedly, if an agent's φ-ing is caused by a desire to make it the case that *p*, it will also be caused by a reason to φ – namely, the fact that φ-ing would make it the case that *p*. But this is precisely the connection that

[21] It's true that an addict who takes drugs because not doing so will cause him harm is less blameworthy than someone who takes drugs just to feel good, even if in both cases the action was wrong all-things-considered. But I agree with Moore (pp. 317–322) that this is a matter of *justification*, rather than excuse. Though both actions were wrong, the former is *less* wrong – the balance of reasons tells more strongly in favour of not taking drugs in the latter case than in the former. All this is perfectly compatible with both addicts being equally responsible for their actions.

appears broken in (some) addicts – their actions are caused by a desire for drugs, but not by any reason to take them. Indeed they may even *recognize* that they lack any reason to do what they're doing, but still find themselves choosing to do it regardless. My view can thus capture a sense in which such people are less than fully responsible for their actions.

For our second illustrative example, let us consider a case with the opposite normative valence. Suppose a young child succeeds in resisting the temptation of a marshmallow now on the promise of two marshmallows later. Let's stipulate that what the child most desires is indeed to eat two marshmallows later. On Moore's view, the child is fully responsible for their action, since (you'll recall) the possibility that an agent may be less than fully responsible for their action only arises on Moore's view in those cases where the agent either fails to choose in line with their strongest desire, or fails to do what they choose to do. By contrast, my view allows for the possibility that the child is less than fully responsible, and thus less than fully credit-worthy for their choice. Suppose for example that their decision not to eat the marshmallow was *partially* caused by the reasons to do so, but only *together* with certain fortuitous, non-reason-conferring features of the child's environment – the time of day, the absence of any distractions, the colour of the marshmallow – but for which they would have given in to temptation.[22] Such a child is only partially responsible for their action, on my view; though they *happened* to do what they had most reason to do on this occasion, they were only partially sensitive to those reasons, and so their action represents a mere partial rational success.

It might help in further clarifying my view to explain how it differs from two versions of libertarianism that Moore discusses (and to which he gives fairly short shrift) in *Mechanical Choices*. First, consider what he calls 'patchy libertarianism', according to which choices can be "'sort of' free and 'sort of' not free, in varying degrees", in virtue of their causes being "weak or partial", thus "leav[ing] room for some freedom, some power, some ability to do otherwise, and thus, some responsibility" (p.270). Here I agree with Moore's endorsement of Strawson's claim that "[w]hatever sense of 'determined' is required for stating the thesis of determinism, it can scarcely be such as to allow of compromise, borderline-style answers to the question, 'Is this bit of behaviour determined or isn't it?'" (Strawson 1974 [2008]: 21). On the contribution account, however, an agent is partially responsible for their action in virtue of it being partially caused *by responsibility-grounding factors*, not partially caused *simpliciter*. Though causation is all-or-nothing, *contributions* to causings can be partial, weak or strong. As Moore himself notes, though it wouldn't make sense to say that an event is, say, 40% caused, "[i]t makes perfectly good sense to say that one factor was 40% the cause of some event, while other factors were 60% of the cause of that event" (p.271) – and this is all the contribution account requires.

Second, consider what Moore calls 'selective libertarianism', which "concedes that all human choices are fully caused, but selects out some causes but not others as excusing" (p.273). In response, Moore insists that "'A cause is a cause,' that is,

[22] A crucial question that arises in this context is whether a particular non-reason-conferring factor should be thought of as part of what *caused* an action or as a mere *background condition*, for the purposes of attributing responsibility; see Kaiserman (2021: § 7) for more on this.

for its challenge to responsibility it cannot matter what kind of a cause sufficiently determines behavior so long as it does at least that" (ibid.). I agree that there is nothing *intrinsically* excusing about some causes of behaviour (poverty, mental illness) as opposed to others (boredom, reading too much Nietzsche as a teenager). But that doesn't mean that some causal explanations of an agent's action can't provide useful information about how responsible they are for the action, *by* providing information on how much reasons (and their absences) contributed to bringing it about. The fact that an agent's action was partially caused by reading Nietzsche twenty years ago tells us precisely nothing about how responsible they are, for example, since it is transparently compatible with the action also having been fully caused by reasons (after all, reading Nietzsche might simply have exposed the agent to reasons to do something that they wouldn't otherwise have been exposed to). By contrast, suppose we learn that an agent's copious writing activity was substantially caused by a metastatic brain tumour (Imamura et al. 1992), or that their decision to help a passerby was substantially caused by their having found a dime in a phone box shortly beforehand (Isen and Levin 1972).[23] These explanations are much harder to reconcile with the claim that the agent is fully responsible for their action; not, to repeat, because of anything to do with the causes themselves, but because of what such explanations imply about how causally sensitive to reasons they were.

Although we clearly disagree on several things, there are aspects of the view I have described in this section to which I think Moore should be sympathetic. Firstly, the contribution account fully embraces a refrain that runs throughout *Mechanical Choices*: that acting freely is perfectly compatible with, and indeed entails, the fact that one's action was caused by prior states of affairs. Indeed, in Moore's explanation of why one is responsible in "the normal case where…one does what one most wants to do" (p.328), modal considerations are nowhere to be found. What makes such cases "paradigms of responsible agency" (p.329), for Moore, is clearly their causal history, not what the agent does in non-actual worlds. Secondly, the idea that causal contribution comes in degrees is one Moore himself has done much to champion.[24] Moore has used this fact in previous work to make sense of degrees of responsibility for *outcomes*; it would be a natural move for him to apply the same approach to the causal facts *upstream* of our decisions and actions, as well as downstream of them.[25]

5 Conclusion

Mechanical Choices is a monumental book, rich with ideas. There is much in it I agree with – the commitment to compatibilism, the rejection of epiphenomenalism, and the rallying cry for the mind sciences to be "the helper rather than the challenger of criminal law" (p.477). Inevitably, I have chosen to focus on an issue about

[23] I discuss the significance of the 'situationist' literature for responsibility in Kaiserman (2021: § 7).

[24] See especially Moore (2009); for discussion see Beebee and Kaiserman (2020: 370–71).

[25] That said, there are elements of my preferred version of the contribution account to which Moore will probably be less sympathetic, particularly my appeal to absence causation (Moore (2009); for recent discussion, see Walen (2022)).

which our views diverge – how to understand the (pre-theoretically plausible) idea that responsibility comes in degrees. According to Moore, degrees of responsibility should be understood in terms of the scalarity of certain capacities, which in turn should be understood in modal terms, as the strength of the antecedent needed to make certain counterfactuals true.[26] I have argued that this makes Moore's view susceptible to familiar counterexamples involving masks and finks. In its place, I have suggested a view on which one's degree of responsibility for one's action depends on the degree to which reasons and their absences contributed to bringing the action about. Interestingly, the view combines two important Moorean ideas: the idea that causal contribution comes in degrees and the idea that acting freely is compatible with (and indeed entails) the fact that one's action was caused by prior states of affairs. But unlike Moore's, it is a view on which responsibility has nothing directly to do with the ability to act, or choose, otherwise.

Acknowledgements Many thanks to Alec Walen and the participants of a workshop in Rutgers for valuable feedback on earlier versions of this paper.

References

Beebee, H. and Kaiserman, A. (2020). Causal Contribution in War. Journal of Applied Philosophy *37*(3), 364–377.

Björnsson, G. (2017). Explaining (Away) the Epistemic Condition on Moral Responsibility. In P. Robichaud and J. W. Wieland (eds.), *Responsibility: The Epistemic Condition*. Oxford: Oxford University Press.

Choi, S. (2012). Intrinsic Finks and Dispositional/Categorical Distinction. Noûs *46*(2), 289–325.

Clarke, R. (2008). Intrinsic Finks. Philosophical Quarterly *58*(232), 512–518.

Coates, D. J. and Swenson, P. (2013). Reasons-Responsiveness and Degrees of Responsibility. Philosophical Studies *165*(2), 629–645.

Cohen, D. and Handfield, T. (2007). Finking Frankfurt. Philosophical Studies *135*(3), 363–374.

Everett, A. (2009). Intrinsic Finks, Masks, and Mimics. Erkenntnis *71*(2), 191–203.

Frankfurt, H. (1969). Alternate Possibilities and Moral Responsibility. Journal of Philosophy *66*(23), 829–839.

Hale, M. (1736 [1847]). *Historia Placitorum Coronae: The History of Pleas of the Crown*. Philadelphia, PA: Robert H. Small.

Holton, R. and Berridge, K. (2013). Addiction Between Compulsion and Choice. In N. Levy (ed.), *Addiction and Self-Control: Perspectives from Philosophy, Psychology and Neuroscience*. Oxford: Oxford University Press.

Imamura, T., Yamadori, A. and Tsuburaya, K. (1992). Hypergraphia Associated with a Brain Tumour of the Right Cerebral Hemisphere. Journal of Neurology, Neurosurgery, and Psychiatry 55(1), 25–27.

[26] Or, alternatively, in terms of the closeness of certain possible worlds.

Isen, A. M. and Levin, P. F. (1972). Effect of Feeling Good on Helping: Cookies and Kindness. Journal of Personality and Social Psychology *21*, 384–388.
Kaiserman, A. (2016). Causal Contribution. Proceedings of the Aristotelian Society *116*(3), 387–394.
Kaiserman, A. (2017a). Partial Liability. Legal Theory *23*(1), 1–26.
Kaiserman, A. (2017b). Necessary Connections in Context. Erkenntnis *82*(1), 45–64.
Kaiserman, A. (2018). 'More of a Cause': Recent Work on Degrees of Causation and Responsibility. Philosophy Compass *13*(7), e1249.
Kaiserman, A. (2021). Reasons-Sensitivity and Degrees of Free Will. Philosophy and Phenomenological Research *103*(3), 687–709.
Kaiserman, A. (2022). Alternative Possibilities in Context. Inquiry *65*(10), 1308–1324.
Lewis, D. K. (1997). Finkish Dispositions. Philosophical Quarterly *47*(187), 143–158.
Mackenzie, G. (1678). *The Laws and Customs of Scotland*. Edinburgh: James Glen.
Moore, G. E. (1912). *Ethics*. Oxford: Oxford University Press.
Moore, M. (1998). Liberty and Drugs. In P. De Grieff (ed.), *Drugs and the Limits of Liberalism: Moral and Legal Issues*. Ithica: Cornell University Press.
Moore, M. (2009). *Causation and Responsibility: An Essay in Law, Morals and Metaphysics*. Oxford: Oxford University Press.
Morse, S. (2003). Diminished Rationality, Diminished Responsibility. Ohio State Journal of Criminal Law *1*, 289–308.
Sartorio, C. (2016). *Causation and Free Will*. Oxford: Oxford University Press.
Sripada, C. (2016). Self-Expression: A Deep Self Theory of Moral Responsibility. Philosophical Studies *173*(5), 1203–1232.
Stephen, J. F. (1883 [2014]). *A History of the Criminal Law of England*. Cambridge: Cambridge University Press.
Strawson, P. F. (1974 [2008]). *Freedom and Resentment and Other Essays*. London: Routledge.
Tugby, M. (2016). On the Reality of Intrinsically Finkable Dispositions. Philosophia *44*(2), 623–631.
Vetter, B. (2015). *Potentiality: From Dispositions to Modality*. Oxford: Oxford University Press.
Vetter, B. (2019). Are Abilities Dispositions? Synthese *196*, 201–220.
Vihvelin, K. (2004). Free Will Demystified: A Dispositional Account. Philosophical Topics *32*(1/2), 427–450.
Walen, A. (2022). More Contra Moore on Absences as Causes. Criminal Law Bulletin *58*, 375–385.
Walker, N. (1968). *Crime and Insanity in England, Volume 1: The Historical Perspective*. Edinburgh: Edinburgh University Press.
Williamson, T. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.