THE “CAN’T/WON’T” DISTINCTION AND THE NATURE OF VOLITIONAL EXCUSE

Michael S. Moore†

NORTHWESTERN UNIVERSITY LEGAL THEORY WORKSHOP
Northwestern Law School
Chicago, Illinois
October 1, 2013

† Walgreen University Chair, Professor of Law, Professor of Philosophy, Professor in the Center for Advanced Study, Co-Director of the Program in Law and Philosophy, University of Illinois. First given to the Conference on Crime, Punishment, and Responsibility: the Legal Philosophy of Antony Duff, Department of Philosophy, Stirling University, Stirling, Scotland; then to the Conference on Free Will, Department of Philosophy, Tufts University, Medford, Mass. My thanks go to the participants at both discussions, and particularly to Dan Dennett, Antony Duff, Doug Husak, Herbert Morris, Kimberly Ferzan, and Suzanne Uniacke, for their many helpful comments.
I. THE DEPENDANCE OF VOLITIONAL EXCUSE ON THE CAN’T/WON’T DISTINCTION

A. The Domain of Excuse

Aristotle famously divided up the excuses from responsibility between cognitive excuses such as ignorance or mistake, on the one hand, and “conative” or “volitional” excuses such as duress, on the other.\(^1\) As to the latter, Aristotle told us that such excuse exists “when one does what he ought not under pressure which overstrains human nature and which no one could withstand.”\(^2\) Moral philosophers and criminal law theorists ever since have struggled to make sense of this second kind of excuse. I shall continue that effort in the present paper.

We first need to bound the topic. Excuses are to be distinguished from that other dominant mode of exculpation, justifications, along familiar lines. For actions that are justified, there is nothing to excuse because there is no wrongful action even when acts of the type in question are prima facie wrong. The domain of excuse is entered only when actions are wrong to do; excuse is a mode of defeating blameworthiness for wrongful actions. Excuses do this by defeating culpability, which along with wrongdoing is necessary for responsibility (blameworthiness) for some untoward result.\(^3\)

Likewise, excuses are to be distinguished from those conditions that rule out there being an act for which one is prima facie responsible. If one’s body causes harm while one is unconscious due to shock, or is asleep, hypnotized, in a post-hypnotic state, in a hypoglycemic episode, or performs a reflex movement or a movement during an epileptic seizure, one does not excuse the “actor” in such conditions. Rather, these cases instance ways in which the bodily motions of a person can cause harm to another without those bodily motions constituting an act of that person.\(^4\) In such cases there is nothing to be excused because one has done no wrongful action.

Thirdly, the volitional branch of excuse is not to be confused with the ways in which an act of a person may not be intended or intentional. Yielding to a threat, a craving, or a passion is not to be confused with doing some wrongful action by mistake, by accident, or in ignorance of

---

\(^1\) Aristotle, *Nicomachean Ethics*, Book III, ch. 1
\(^2\) Id.,
the kind of act one is doing. Unlike some of the cognitive excuses, volitional excuses come into play for actions that are intentionally done by an accused. Although often compelled people say that “they had no choice,” they don’t mean that literally; for a compelled choice is still a choice, and the action chosen is both intended and done intentionally. If there is a volitional excuse in play, it does not lie in the negation of choice, knowledge, or intention. Volitional excuses, unlike some of the cognitive excuses, only come into play for actors who are admittedly prima facie culpable (because knowing or intentional) in the doing of their admittedly wrongful actions.

Putting these three points together, we should see volitional excuses as occupying part of the fourth quadrant in the four square box of responsibility.\(^5\)

<table>
<thead>
<tr>
<th>WHAT IS NEEDED FOR PRIMA FACIE</th>
<th>WHAT IN ADDITION IS NEEDED FOR ALL OUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>WRONGDOING</td>
<td></td>
</tr>
<tr>
<td>1a. Voluntary act which causes a prohibited harm;</td>
<td>3. Without Justification</td>
</tr>
<tr>
<td>or 1b. Willed omissions to prevent a harm which one has the duty and the ability to prevent.</td>
<td></td>
</tr>
<tr>
<td>CULPABILITY</td>
<td></td>
</tr>
<tr>
<td>2. While intending, foreseeing, or consciously risking a harm of that type.</td>
<td>4. Without Excuse</td>
</tr>
</tbody>
</table>

One is blameworthy for a harm when one is a culpable wrongdoer, and excuses are but one of the ways in which one may not be such a culpable wrongdoer.

I thus put aside those cases of duress, necessity, and other nominally volitional excuse defenses, when legal doctrines allocate such defenses to non-excuse categories. Necessity and duress, for example, are often so defined in the criminal law that they operate exclusively as justifications and not as excuses. Under the Model Penal Code necessity is treated exclusively as a matter of justification.\(^6\) Likewise, prevailing American common law implicitly treats duress exclusively as a justification by defining: (1) the harm allowably threatened (to constitute a


\(^6\) Model Penal Code § 3.02.
defense) so severely that it excludes all but death or grievous bodily injury; and (2) the harm one can cause to alleviate the harm threatened sufficiently less severely that it excludes death; and (3) the degree of necessity of doing the latter in order to avoid the former quite stringently; with the result that only justified actions can count as actions “excused” because of duress.7

Similarly, some courts and commentators urge that a compelled act is really no act of the defendant at all, or at least that it is no voluntary act of the accused, it being seen as on a par with hypnotized or reflex “acts.” In a leading English decision, for example, the Court of Appeals characterized the defense of duress as existing “if the will of the accused [was] overborne…so the commission of the alleged offense was no longer the voluntary act of the accused.”8 Others regard compelled actions as acts where “individuals merely react rather than choose to do wrong,”9 and without choice there can be no intention or other forms of culpability.

Defenses along any of these three lines are not our interest, for all of them involve exculpating factors that are not excuses. Wanted is a restriction to those instances of legal defenses, and the moral features that ground them, that occupy the domain of excuse.

B. The Domain of Volitional Excuses

Within the category of excuse we also need to narrow our focus to volitional excuses. One distinction we need is Aristotle’s, between cognitive versus conational excuses. Ignorance of what one is doing or causing, ignorance (in some cases) that it is wrong or prohibited, or mistakes about either of these matters, can thus be put aside for present purposes.

More troublesome is a second distinction that we seemingly need, one drawn within the “compulsion” or “coercion” excuses. This is Herbert Hart’s distinction between compulsions that operate through incapacity and those that operate through absence of fair opportunity.10 As a first cut at the distinction, incapacities can be seen as equipment failures: the accused couldn’t have done otherwise because he didn’t (at that time at least) have the equipment to do better. Whereas absence of opportunity is an excuse available to those with perfectly functioning equipment, just not a fair opportunity to use it.11 To see the distinction, think of two ways of looking at addiction, assuming arguendo that addictions excuse: does addiction excuse because the craving for drink, drug, food, sex, etc. incapacitates the will from its usual ability to control

---

8 Regina v. Hudson, [1971]-2 All E.R. 244 (Crim. App.).
11 I put the distinction this way in Moore, Placing Blame, p. 554.
such desires? Or does addiction excuse because the addict faces a cost non-addicts do not have
to face when they decide not to use drugs, namely, the cost of withdrawal? The latter sort of
addict may foresee the cost of withdrawal coolly and without incapacitation of his will, but
claims excuse nonetheless because he faces opportunity costs (not to do wrong) that non-
addicted persons do not face.

To focus on volitional excuses, one might be tempted to completely put aside any “unfair
absence of opportunity” versions of defenses like duress, necessity, addiction, and the like. This
would seemingly allow us to focus on actors whose excuses are framed in terms of not being
able to do better, rather than framed in terms of it being unfair to ask them to do better in light of
their unusually high opportunity costs. Yet I now think that the incapacity and lack of
opportunity branches of defense like duress, addiction, and the like, are too closely related for
such a set-aside to be possible.

Let me begin with a criticism of lack of opportunity as a separate kind of excuse, a
criticism pressed against my earlier views on this by Antony Duff and Victor Tadros. To
answer the criticism adequately will show why lack of capacity versus lack of opportunity cannot
be so easily separated into two kinds of non-cognitive excuses.

I shall put Duff’s and Tadros’ criticism my own way. Lack of fair opportunity to avoid
doing a wrong action has only three plausible instances, it seems to me. The first is where
something worse will happen if one does not do the wrong in question; one’s opportunity to
avoid doing a certain wrong then comes at the unacceptable cost that one would be allowing an
even greater wrong to take place. The second is where one’s body becomes an instrument of
causing harm through no willing of one’s own, as where a third person or a gust of wind grabs
your arm and causes it to strike some victim; in these cases too you have no opportunity to use
your in-good-working-order equipment to refrain from doing wrong. Here too you had no fair
opportunity. The third is where one is ignorant or mistaken about the need not to do some act,
because one is ignorant or mistaken about the aspects of that act that will make it wrong to do.
So long as one’s ignorance or mistakes are not themselves culpable in their origins, here too one
lacks the opportunity to use one’s abilities to prevent wrongdoing. Yet, these three instances of
lack of fair opportunity are already taken care of by: (1) justification; (2) the requirement of

---

action; or (3) the requirement that there be culpable intention or knowledge. None of these instances of lack of fair opportunity require that there be an excuse of lack of fair opportunity. Indeed, if a forbidden act is unjustified (and thus wrong), and if it is done intentionally by an actor who has no volitional excuse (i.e., he could have refrained from doing it but went ahead anyway) – why would anyone ever be tempted to excuse such a wrongdoer? He intentionally did some wrongful action that he didn’t have to do; how can that be blameless?

Duff and Tadros plainly have their finger on something, do they not? If one explores the objection further, that only seems to confirm its potency. If one tries to think of a fourth kind of loss of opportunity case (where exculpation is plausible but where that would not be due to presence of justification or of incapacity, or due to lack of act or intent), one is apt to think of threats to one’s children or others who are near and dear. When an evil threatener threatens a parent with the death or serious or painful injury of his child unless the parent does some act producing some greater evil, such as revelation of important state secrets, excuse does seem plausible. Now stipulate that the threatened parent is cooler in his emotions than even Kant could imagine; he coolly but firmly chooses not to allow his child to be injured. Is such a parent still blameless, even though he is in no way incapacitated in his reasoning abilities by fear for his child’s safety or by anger at his child being so threatened? My own judgment (at least in some of such cases) is that the parent is blameless. Yet this is not because the parent is excused by the unfair lessening of his opportunity to keep his child safe. Rather, it is because parents have agent-relative obligations to safeguard their children even when the harm threatened to their children is much less (on an agent-neutral basis) that the harm such parents must do to avert it.14 So these are cases of justification after all, and there is no need for an excuse of lack of fair opportunity to account for exculpation in such cases.

Here is how I now see the “lack of fair opportunity” excuse. It is not a separate sort of excuse, discontinuous from the incapacity sort of excuse. Rather, there is only one excuse here, but it is an excuse where both the degree to which the actor is incapacitated and the degree to which he lacks fair opportunities to avoid doing evil, each play their part. Cases where the degree of incapacitation is not so great but where the opportunity costs of not doing wrong are

high (but not so high that they justify doing the wrong), we might well call “lack of opportunity” cases; in contrast to cases where the degree of incapacitation is high but where the opportunity costs of not doing wrong are not high (or if they are high, that is not salient) – these we might think of as “incapacity” cases. In reality all cases will require both features if they are to be plausibly cases of excuse. Labeling them as one or the other reflects nothing more than the salience of one factor over another.

My own earlier examples pointed to this conclusion, although I didn’t see it. One is that of the saint-hating bank robber, A.\(^{15}\) A hates a saintly person, S, and is delighted that S is dying of a rare disease. B threatens A with B giving S a life-saving antidote unless A drives a get-away car for B in B’s contemplated bank robbery. A is distressed beyond description by the thought of S surviving, so to drive the image of S’s saintly presence on this earth out of his obsessed mind, A reluctantly agrees to drive the get away car, and does so. No one thinks A’s incapacitation excuses. This, because he had a \textit{fair} opportunity not to be a bank robber; his only “cost” was the survival of a saint, not objectively a cost at all but a benefit.

One of my original conclusions from consideration of this kind of case was correct; incapacitation by itself is insufficient to excuse; lack of fair opportunity is also necessary. But my other conclusion from such cases was that lack of such opportunity was also sufficient for excuse, when there is no incapacitation, and this is incorrect. Imagine someone whose opportunities not to do some wrong are severely diminished; but that his perception of the diminished opportunities in no way incapacitates him. The diminished opportunities may show that the wrong on this occasion is not so great as it would have been had he had greater opportunities not to do it, because some good will come out of the doing of such a wrong on this occasion. The \textit{net} wrong, we might say, is small in such cases. Thus, what it takes to excuse such wrong will be less (as Austin famously remarked, it takes less to excuse stepping on a snail than it does to excuse stepping on a baby).\(^{16}\) Yet notice that nothing in the lack of fair opportunity story does any excusing. Lessening the wrong done means we need less excuse, but we still need some. Without some incapacitation, even a small wrong will not be excused by mere diminishment of opportunity (at any level less than that of justification).

\(^{15}\) Moore, \textit{Placing Blame}, p. 556.
So we need both some degree of incapacitation and some degree of lack of opportunity to have volitional excuse.\(^\text{17}\) Which is why we cannot put aside the “lack of fair opportunity cases” from the domain of cases we want to consider. For the cases that are so labelled are only incapacity excuse cases with not so serious incapacitation doing the excusing because decisions where the opportunity costs not to do wrong were quite high.

It may seem that, apart from these moral considerations militating against hiving off some separate lack of opportunity excuse, there are conceptual reasons not to do so as well. Such conceptual reasons would have to do with alleged difficulties in distinguishing lack of opportunity from lack of capacity. Yet as a first cut this distinction seems clear enough, for lack of capacity has to do with defective equipment and lack of opportunity, with the fair chance to use non-defective equipment. When we are speaking about general capacities, this is an unproblematic distinction.

Capacities can be general or particular.\(^\text{18}\) “I \textit{can} run a mile in under five minutes” refers to my general ability to do a class of actions. But capacities can also be particular, as in: “I could have run the mile in the track meet yesterday in under five minutes.” The latter does not simply mean that I was possessed of the general abilities (of mile-running-under-five minutes) over some swatch of time or even at race time (t) yesterday;\(^\text{19}\) I do need such general abilities at t if indeed I could have run a mile in under five minutes at t. But I might have such general abilities

\(^{17}\) This lessens but does not eliminate the gap between my views on volitional excuse, and the current, “Aristotelean views” of Antony Duff. See Duff, “Virtue, Vice, and Criminal Liability: Do We Want an Aristotelian Criminal Law?,” \textit{Buffalo Criminal Law Review}, Vol. 6 (2002), pp. 147-184; Duff, “The Virtues and Vices of Virtue Jurisprudence,” in T. Chappell, ed., \textit{Values and Virtues} (Oxford: Oxford University Press, 2007). For now both Duff and I have an incapacity part to the excuse, and a “moralized” part, even though each of these perhaps differ between us. Duff now distinguishes two kinds or levels of volitional excuse: (1) those where there is severe incapacitation, defined as instances where the actor is “so terrified…that he is no longer capable of the kind of practical reasoning that would ground a rational decision…,” excusing unless “a reasonable person would not have lost control of herself as this defendant did…” (“Virtue, Vice, and Criminal Liability,” p. 178); and (2) those where the incapacitation of reason is less severe, where the emotion felt by the actor is “apt to destabilize” one’s reason or “disturb” one’s rational deliberation, excusing so long as the emotion is a “reasonable” one, “reasonably motivating” an actor to do what this actor did, and does not constitute vice (“The Virtues and Vices of Virtue Jurisprudence”). One suspects there is still more of a character assessment in Duff’s moralized conditions of reasonableness than in my requirement of a “good-but-not-good-enough-for-justification” sense of diminished opportunity; but one would have to work out the details to see. For a third variation also finding pure incapacitation “unmoralized” to be insufficient as excuse, see Claire Finkelstein, “Duress – A Philosophical Account of the Defense in Law,” \textit{Arizona Law Review}, Vol. 37 (1995), pp. 251-283.


\(^{19}\) Honore, “Can and Can’t.”
at t, and still lack the capacity to have run that mile in under five minutes on that occasion. General abilities do not guarantee success on every occasion, and for some failures by those possessed of general abilities it may be true that the actor could not have done better on that occasion.

As we shall see, the capacities in which we are interested for purposes of volitional excuse are ultimately capacities on a particular occasion. We will want to know of a given actor who has done a wrong on a particular occasion, could he have done otherwise on that occasion? It is true that those who lack the capacity (general) to do better in these types of situations, will not have the capacity (particular) needed for responsibility. But actors possessed of the capacity (general) might still lack the capacity (particular) needed for responsibility, so that latter question will still need answering after we have settled the question of the actor’s capacities (general).

Now we can see the challenge of distinguishing lack of opportunity from lack of capacity. All is well with the distinction at the level of capacities (general), so that if the actor lacked the capacity (particular) not to do some wrong that he in fact did because he lacked the “equipment” (capacity-general) ever to do better in such kinds of situations, then we can meaningfully distinguish such lack of capacity (particular) from lack of opportunity. But for cases where the actor could not have done other than the act he did on some particular occasion – where he lacked the capacity (particular) to do better, in other words – even though possessed of the “right stuff” (capacity-general), it may be difficult to discern some separate lack of opportunity. When equipment is fine but the actor fails nonetheless in a way that seems excusing, did he fail for lack of capacity (particular) or lack of opportunity?

Suppose one is being blamed for forgetting an important appointment. Suppose further we are confident that the actor could not have remembered this appointment on this occasion, and yet we are also confident that his “memory equipment” (hypothalamus etc.) is in good working order so that he had the capacity (general) to remember such things. Can we now parse, “he could not have remembered this appointment at t,” into two possibilities, incapacity versus a lack of opportunity? It is not just that we can’t tell which it is; it is not obvious that it is one rather than the other.

In such examples, notice, we cannot use the intuitive basis we relied on before to distinguish two kinds of heroin addicts seeking to excuse their acts of using heroin. There, one can distinguish the addict whose judgment-making processes are unhinged by their craving for
heroin (incapacity), from the addicts whose judgment-making processes are not so disturbed but who judge (wrongly) that the costs of withdrawal are too high a price to pay for doing the right thing, i.e., not taking the drugs (lessened opportunity). For unintentional omissions such as not remembering an appointment, there is no disturbed versus wrong judgment distinction to be made.

I conclude that the incapacity/lack of opportunity distinction only makes obvious sense for some but not all of the volitional excuse cases we want to consider. For this conceptual reason too, in addition to the moral reasons given earlier, we may do well not to think we can put aside some separate, lessened opportunity version of volitional excuse.

C. What Legal Defenses Plausibly Involve Volitional Excuses?

The range of plausible volitional excuses is broad. Volitional excuse figures in the standard legal characterization of various instances of the defenses of necessity, duress, addiction, provocation, innocent aggressor self-defense, excessive force self-defense, involuntary intoxication, immaturity, insanity, diminished capacity, and diminished responsibility. A traditional organization of the volition based instances of these defenses is in terms of the source of the factor that is said to compel or coerce an actor into intentionally doing some wrongful action. Externally sourced compulsions include duress, where the threat is human sourced, or natural necessity, where the threat is one imposed by natural features. Internally sourced compulsions include provocation, where the actor’s own emotion (anger) is said to compel him to act; addiction, where the actor’s own craving for drink or drugs is said to compel his behavior; self-defense against innocent aggressors, where the actor’s fear or instinct of self-preservation is said to make it difficult not to hurt or kill an innocent; excessive force used in self-defense, where one’s arousal at being attacked carries over into wrongful retaliation.

Yet this taxonomy (between internal and external sources of compulsion) is superficial. For in the “external” cases as much as the “internal cases,” the same kind of incapacitation of will seems to be the driving force of excuse. When threats from other people (duress) or from nature (necessity) seem to excuse, it is the fear, self-preservative instinct, craving, or other “internal” factor through which the threat operates that is doing the incapacitation. The external source of the emotion that incapacitates drops away as non-essential.

---

The internal versus external source distinction is in any event not to be found in the “volitional prongs” of the status excuses of insanity, diminished capacity, diminished responsibility, involuntary intoxication, and immaturity. Status excuses are so-called because they are based on general characteristics of an accused (his “status”), not on the particular features present on the very occasion on which he did the wrongful act or made the culpable choice. As with the non-status defenses, standard legal statements of some of these status defenses are based on volitional rather than cognitive excuse. In these cases (the volitional ones), such conditions as mental disease, youth, retardation, other mental defects, and intoxication, all are thought to be excusing because all are thought to give rise to “irresistible impulses,” “inabilities to conform one’s behavior to the requirements of law,” lack of “substantial capacity to do what he knows the law commands,” etc. What is said to be incapacitating in these defenses is not always based on the strength or the intensity of an emotion felt on a particular occasion, but rather, the lack of the equipment (“executive functioning”) with which to control one’s “impulses.” In such cases the incapacity to have acted otherwise seems based on an incapacity (general).

D. The Underlying Principle of Volitional Excuse: The Actor Could Not Have Done Otherwise

I take the volitional excuse prongs of all of these legal defenses, and all of the moral excuses that underlie them, to exemplify but a single principle of exculpation. This principle is sometimes put in terms of one being “deprived of his free will” by the coerced or compelled actor; sometimes as a “loss of power to choose” the good and the lawful because of a will that is “overborne;” sometimes as an “impairment of the actor’s ability to control his conduct.” Inability, lack of power, lack of freedom, are all ways of referring to actors who intentionally do wrongful actions and yet who couldn’t have done otherwise, whereas wrongful acts intentionally done by those possessed of the ability to control their conduct, those who had the power to choose the good and the lawful, and those who were free to do as they ought, are said to be done by actors who could have done otherwise but simply wouldn’t do so. Thus the crucial issue for

---

21 I taxonomize the excuses this way in *Placing Blame*, pp. 483–485.
22 My own view is that the standard legal statements of status excuses like insanity are quite wrong in their reliance on volitional incapacities, whether particular or general. See, most recently, Moore, “The Quest for a Responsible Responsibility Test: Norwegian Insanity Law After Breivik,” forthcoming.
all such defenses is to distinguish between two classes of defendants who in fact didn’t refrain from doing illegal and immoral actions: those who are excused because they couldn’t do better, and those not excused because they just wouldn’t do better.

The trick is to articulate senses of *power, ability, freedom*, and what one *can* do, that give content to the principle underlying all volitional excuse.²⁶ There are two extremes to be avoided here, one of which eliminates all responsibility (because we are all unable to do other than we in fact did), and the other of which confers responsibility on all wrongful and intentional actors (because we are always able to do other than what we intentionally did). Call these the “ultra-liberal” and “ultra-conservative” extremes, respectively.

The ultra-liberal’s conclusion stems from two beliefs. One is that we *can* only do—equivalently, we have the power, ability, or freedom to do—only what it is *possible* to do, and for some act or choice of ours to be possible requires that it not be caused by factors themselves unchosen (or otherwise outside the actor’s control). Call this first belief a belief in the incompatibilist sense of “can,” so-called because having an ability to do otherwise in this sense is incompatible with causation of choice by factors themselves unchosen. The second belief is that of causal determination of all human choice and action—“determinism,” for short.²⁷ These two beliefs generate the skeptical conclusion that none of us have the ability to do other than we actually did do on some occasion, and thus that none of us are responsible for anything. The ultra-liberal’s reasoning goes like this: if by causation we mean something strong like sufficient conditions, then what is caused “had to happen” given what went before. If something had to happen, then there was no possibility of anything else happening. This applies to human choices and actions: if they were caused in this strong sense, then there was no ability to have chosen or acted other than we did on some occasion. Since it is very plausible that all human choices are as caused in this sense as are any other natural events, there never is such power, and there thus is never any responsibility.

The ultra-conservative’s conclusion is almost as devastating of our intuitive views. There are several variations of this view. The main variation of this view uses quite different senses of

---

²⁶ In what follows I will ignore the (sometimes intuitive, sometimes not, but always subtle) distinctions between these terms that philosophers have stipulated.
²⁷ I also prescind from the philosophical thickets of either defining or defending determinism in its usual definition. Only needed by the ultra-liberal is the thesis that all human choices and actions are caused, *in whatever sense and to whatever extend of causation as is true of natural events like earthquakes, accidental falls, etc.*
“can,” “ability,” “power,” and “free.” Beginning with Hume, the central idea is that we are at liberty – free – whenever our choices (or intentions) cause the actions chosen (intended). We have the power needed for responsibility, the ability, the free will, whenever we cause what we choose to cause. This is usually termed the compatibilist sense of these terms, because the causation of actions by our choices to do those very actions is quite compatible with such choices themselves being caused by factors outside our control. Being a causer in no way require that one be an uncaused causer.

In this sense of “can,” the principle asking whether one could have done otherwise is always elliptical for, “could have done otherwise if he had chosen to do otherwise.” The principle, in other words, holds us responsible wherever we have the power to effectuate our choices. It is silent whether our choices are sufficiently caused by factors themselves unchosen. The principle is silent in the sense that it does not require contra-causal freedom for our choices in order for us to be responsible; caused choices may still cause the actions chosen, and thus satisfy this version of the principle of responsibility.

Someone whose arms are being moved by another in ways causing harms to others, or who is similarly restrained by nature, lacks the power to effectuate his choice to do other than cause such harm; such a person is thus excused by the ultra-conservative’s version of the principle. But then, such a person is not even acting. Those who do perform actions, and do so intentionally, can have no excuse under the Hume/Moore reading of the principle as thus far construed. This is indeed the ultra-conservative’s conclusion. If I yield to a severe threat by doing some harmful action A wanted by my threatener, I did have the power to effectuate my choice not to do A. True, if I didn’t do A I would suffer the adverse consequences of the threat. But if I had chosen not to do A, I would not have done A, and this is enough to make me responsible for doing A under the ultra-conservative interpretation of the principle. This illustrates how no volitional excuse can exist under this version of the principle as thus far construed. For that the accused did what he intended to do in response to the intention to do it, is sufficient for responsibility under the principle, no matter how hard it may have been for him not

28 David Hume, “Of Liberty and Necessity,” Enquiry Concerning Human Understanding (1748), VIII.
29 I put aside the arguments of those who think that there is something unique about persons (or something unique about the agential causings of persons) such that persons can cause things to happen only if persons are themselves uncaused in their causings.
to have chosen as he did and no matter how hard it may have been for him to act in conformity with his choices.

The ultra-liberal and ultra-conservative interpretations of the principle of responsibility are not idealizations held by no one but philosophers. The power of each interpretation has been felt by those who design and operate legal institutions. The ultra-liberal interpretation finds voice every time some new defense is proposed because of discovery of some new cause of certain criminal behaviors.\(^{31}\) Witness the old “XYY defense,” the “PMS defense,” the “rotten social background” defense, the “brainwashing defense,” the “disintegration of self defense,” etc.\(^{32}\) The ultra-conservative interpretation finds voice whenever courts or commentators assure us that successfully executing an intention is all that one need do to be responsible for the results intended, or that difficulties in how such intentions are formed or executed are imaginary, or at least unprovable. Thus we have witnessed the belief that because no one yields to compulsions “with a policeman at their elbow,” no impulse is truly irresistible, only unresisted;\(^{33}\) that loss of control tests are conceptually incoherent and in any event factually unverifiable;\(^{34}\) that addictive cravings never rob one of an ability to stay sober or do what one ought to do, if the person wanted enough to do so.\(^{35}\)

As philosophically venerable, popularly accepted, and legally influential as these two extreme interpretations are, quite obviously neither principle can serve to isolate excuses of compulsion. The first makes such excuse universal; the other eliminates it entirely. Needed is

---

34 Richard Bonnie, “The Moral Basis of the Insanity Defense,” American Bar Association Journal, Vol. 19 (1982), pp. 194-196, at 196 (“There is... no objective basis for distinguishing... between the impulse that was irresistible and the impulse not resisted...”). Incoherence and unverifiability are the two leading reasons leading both the American Psychiatric Association and the American Medical Association to recommend the elimination of any volitional prong to legal insanity tests. American Psychiatric Association Statement on the Insanity Defense (1982). This has influenced some courts in their doing just that. See United States v. Lyons, 731 F.2d 243, 739 F.2d 994 (5th Cir. 1984).
35 Powell v. Texas, 392 U.S. 514, 522 (1968). (Thurgood Marshall questioned whether the alcoholic cannot stay sober, observing that Leroy Powell managed to do so when the costs of not doing so are high enough, such as when Powell needed to appear sober in court.) On this line, see generally Herbert Fingarette, Heavy Drinking. More recently Antonin Scalia has also questioned whether anyone who intentionally acts in response to his own desires, ever is “out of control.” Kansas v. Crane, 534 U.S. 407, 423-424 (2002).
some other interpretation that allows there to be a distinction between what we couldn’t help doing and what we simply wouldn’t refrain from doing.

There are two ways to approach the formulation of some such more moderate principle of responsibility; one could soften the harshness of the ultra conservative interpretation by allowing conditions beyond mere lack of causal power of choice to constitute an inability to do other than we did. This would broaden excusing conditions. Alternatively, one could lessen the universality of excuse generated by the ultra-liberal interpretation by allowing that the contra-causal freedom that it demands for responsibility can be found in some cases of intentional wrongdoing. I have elsewhere explored the latter option, so I shall here only pursue the former.

II. THE FOLK PSYCHOLOGY OF VOLITIONAL EXCUSE

Stephen Morse has long challenged the coherence (as well as the verifiability) of any loss-of-control excuse. He concludes that: “To date, advocates of an independent control test have not demonstrated the ability to identify ‘can’t’ versus ‘won’t.’” Morse’s main reasons for this conclusion are that: “There is no … measure of lack of control nor is there yet an adequate folk psychological process that has been identified as normatively justifiable for legal purposes.” The aim of this part II is to supply a description of the folk psychological processes defining loss-of-control in a morally relevant way. The aim of the next succeeding part, part III, is to operationalize the folk psychological models of loss of control so that one indeed has some measure of loss of control.

A. Introduction: The Psychology of Conflicting Desires in the Normal Case Where There is Responsibility

In order to distinguish cases where one could have acted or chosen to act otherwise from cases where one could not, we need to be clear in our own minds about the mental machinery that is involved, first, in the normal case where one is responsible because one does what one most wants to do, and second, in the less typical case where one is arguably excused because in some sense he couldn’t form or effectuate his will to do what in some sense he wanted to do. Let us see if we can make this more precise.

---

38 Id.
Start with the simplest case first, where one acts in light of a single desire. Our two-thousand year old picture of rational action is drawn in terms of three representational states, desire, belief, and intention. We represent the world as we want it to be, as we believe it is, and as we intend to make it. As “premises” in a “practical syllogism,” these three representational states form an ordered triad:

1. Actor X Desires (state of affairs G)
2a. X Believes (the world is in state S)
and 2b. X Believes (in S, if X does action A, then G)
Therefore: 3. X Intends (A)
Therefore: 4. X performs A

This is of course too simple. Specifically, there is almost always more than one desire that bears on a given action, and conflict between such mental states is a fact of life, in almost all real world situations. Particularly relevant here are conflicts of intentions as well as of desire. There are two kinds of conflict for each of these two kinds of mental states. The least often occurring kind of conflict is where the propositional content of desires or intentions form contradictories: I desire to eat chocolate cake, and I desire that I not eat chocolate cake. More common is conflict, not in the content of desire, but in the non-satisfiability of two or more desires on a given occasion: I desire to eat chocolate cake, and I desire to remain thin, and to achieve one is to frustrate the other.

Desires and intentions differ in the consequences of their conflicts in that it is criticizably irrational either to intend (A) and intend (~A), or to intend (A) and intend (B) while believing that one cannot do both actions A and B. By contrast it is not criticizably irrational to have conflicts of desire. But criticizable irrationality is not to be confused with psychological impossibility; importantly for present purposes, one can have conflicting intentions even though it is irrational to do so.

With this brush cleared, now suppose a normal case of responsible choice where there is a conflict in the agent’s desires. E.g.:

1. X Desires (to remain thin)
2. X Believes (to remain thin he must not eat the chocolate cake in front of him)

---

Yet it is also true of X at the same time both that:

1. X Desires (to eat the chocolate cake)
1." X Desires (to eat the chocolate cake) more than he Desires (to remain thin)

Therefore: 3. X Intends (to eat the chocolate cake)
And: X does eat the chocolate cake.

In normal cases of conflict, where we do what we most want to do, there is no hint of excuse. On the contrary, such cases are paradigms of responsible agency.

So what changes give rise to intuitions of excuse of the kind in which we are here interested? There seem to be two possible kinds of cases, one where the agent is unable to form an intention that resolves the conflict in his desires, and the other where the agent is unable to act on the conflict-resolving intention that he is able to form. Because separating the cases in this way depends on how we conceptualize beliefs, desires, intentions, and strengths of desire, I shall first make clear my stipulations in my use of these terms.

B. Desire, Strongest Desire, and Intention

We need several terminological stipulations in order to be clear in our conceptualizations of the folk psychology of compulsion. One is to accept the distinction between desires and intentions. Although there is a long tradition in philosophical psychology regarding intentions as a kind of “all-out” want, the better view (that I and others have argued for elsewhere) is that intentions execute the motivations that desires give, and in this intentions are functionally (and probably structurally) distinct kinds of mental states from desires. My first stipulation is to so regard them.

Second, I need a broadly motivational reading of “desire.” This is partly to bypass the nuances in ordinary English usage of “wish,” “want,” and “desire.” But even more, this is to encompass all behaviorally-inclining, motivational states under the rubric of “desire.” This will include moral beliefs about what one ought to do, emotional states, those generic inclinations to action sometimes called “impulses,” reflected-upon goals as well as impulsive cravings, etc. – in short, any state that can serve as the motivating state in a belief/desire pair. Donald Davidson

---

termed such a generically motivating state a “pro-attitude,”⁴² and this is the sense in which I intend to use “desire.”

Thirdly, I need to stipulate the sense I intend for “strength” of desire and, by implication, for the “strongest” desire in situations of competition between conflicting desires. The generic, motivational sense I stipulated for “desire” suggests although it does not require the causal sense I will give for “strong” and “strongest” desire. This, because desire in the generically motivating sense covers all entrants that can compete as reasons motivating a rational agent as she chooses between conflicting inclinations. There will thus be no room for a Kant-like distinction between desires in some narrower sense (Kant’s “inclinations”), and some non-desire attitude such as Kant’s, “reverence for the moral law.”⁴³ Nor between desires in some narrow sense, and “reason” in some neo-Humean sense. All the items that compete for dominance in our deliberations about what to do are desires in the generic motivational sense. This makes natural (although it does not compel) the causal sense I shall stipulate about strength of desire: necessarily the strongest desire (amongst some set of such desires possessed by an agent that each bear on some contemplated course of action) is the desire that wins out in situations of known conflict; it is the desire that causes both the act satisfying such desire and the intention (or choice) to do such an act. Strength of desire, accordingly, is degree of causal power over choice and action.⁴⁴

The stipulated sense to strength of desire rules out other senses that in other contexts can be plausible enough. There is the sense equating strength of a desire with various phenomenological features of desire: degree of longing experienced by the subject for the object of desire, for example; or degree of “have-to-have-this” craving, if that is different; vividness of presentation; etc.⁴⁵ There is also a sense equating strength of desire with degree of satisfaction felt by a subject when the desire is satisfied. There is also a sense equating strength of desire with the desire with which the agent most self-identifies. I put aside all such alternative senses of strength of desire simply on grounds of clear exposition. Everything such senses want to add

---

⁴² Davidson, “Actions, Reasons, and Causes,” in his Essays on Actions and Events.
⁴³ Kant, Groundwork.
⁴⁴ I shall modify these definitions of strength of desire, and strongest desire, later on to accommodate the insights of the folk psychology of when strongest desires that ordinarily would win out in the determination of choice and behavior fail to do so because of the aberrational features constituting loss of control.
to strength of desire can be added on later, as attributes of particular kinds of strong desires, if that is needed. So I make do provisionally with a spare, semi-behavioral notion of strongest desire as the desire that wins out in situations of known conflict of desires.

C. Inability to Form the Right Intention

Of the two kinds of cases earlier distinguished, I shall focus here on cases of the first kind, cases of incapacity in the formation (rather than in execution) of an intention. There are three or four possibilities as to how the agent is unable to form a conflict-resolving intention. One is where the agent forms no intention at all (and thus does not form the intention in 3 above). In such cases the will is by-passed; the desire in 1 directly causes the action, with no mediation of choice or intention (although there must be an executory volition for there to be action). A second kind of case is where intention 3 does exist, but so does a conflicting intention, 3:

3. X Intends (not to eat the chocolate cake)

In the latter sort of cases the agent is unable to form a conflict-resolving intention because his two intentions merely replicate the conflict in his two desires.

A third possibility alters the relative strengths of the competing desires, so that 1” (rather than 1 is true:

1.” X Desires (to eat the chocolate cake) less than he Desires (to remain thin)

Despite this, and despite the maintenance of his belief in 2 that his desires are inconsistent, X nonetheless forms the intention in 3 and acts on it.

We have not yet put our finger on what it is about any of these sorts of cases that might incline one towards excuse (assuming arguendo that eating chocolate cake is wrong in these circumstances and thus needs excusing). Take the first variation, where there is no intention so that desire causes action directly. On a Davidsonian view of rational action these might seem especially irrelevant to excuse because on such a Davidsonian view actions just are desire-caused bodily movements, with no need of any mediating intention. Yet even Davidson and Davidsonians conceded the need of “all-out wants” that have as their function the resolution of conflict between component desires, so the by-passing of will cases will exist for them too, even if they are differently labeled.

46 Moore, Act and Crime, ch. 6. Volitions are a kind of intention, but are the executory intentions to move our bodies. They are distinct from the intentions higher up in the hierarchy of our plans that resolve conflicts in our desires.
Yet are we to believe that the will is really by-passed in such cases, so that the gravamen of excuse lies here? Suppose my desire for chocolate cake causes my heart rate to go up (in anticipation of eating some).\textsuperscript{47} Or suppose that that desire causes me to say “cake” rather than “lake” in some sentence;\textsuperscript{48} to dream of chocolate cakes;\textsuperscript{49} to react more quickly in actions saving cakes from destruction than I act to save bags of squash; etc.\textsuperscript{50} All of these are cases where the will is bypassed and the conclusion of non-responsibility is plausible. Are cases of compulsion often or ever like this? I doubt it. Sometimes unconscious nibbling on items like chocolate cake does occur, as psychologists have studied.\textsuperscript{51} But yielding to compulsion is not at all like such cases. In my supposed case of eating chocolate cake, unlike the Freudian and Bargean cases above, there is plainly action: the agent wills the bodily movements that constitute eating the cake. Moreover, the phenomenology of choice and intention is present; he experiences his eating as chosen, not something he is surprised to see himself doing (like unconscious nibblings). So thus far we have seen little to support the no-intention (or “by-passing of the will”) version of volitional excuse.

Yet perhaps I have left out the crucial feature of the by-passing account, and that is that the desire in \textsuperscript{1} is emotionally charged. In cases where excuse is tempting, such desires as \textsuperscript{1} are: experienced as cravings, as in addictions; experienced as fears, as in duress or innocent aggressor self-defense; experienced as anger-driven, as in provocation. The thought might be that it is the emotional nature of such desires that by-passes the will in a way distinct from such by-passing in cases of “automatic actions.”

Notice that this is not the question of whether strong emotions are excusing even when intentionally acted on. I will get to that question too, shortly. Rather, here the question is whether strong emotions bypass the will like automatic actions: they cause without the mediation of intention and choice. An example is that of a prisoner who desperately wants to escape from

\textsuperscript{47} The kind of examples of “mental causation without motivation” in Moore, \textit{Law and Psychiatry}, p. 15.
\textsuperscript{48} The kinds of slips of the tongue Freud dubbed the “psychopathology of everyday life.” \textit{Id}, pp. 311-312.
\textsuperscript{49} \textit{Id.}, ch. 8.
\textsuperscript{51} Daniel Wegner details the automatisms involved in “ideomotor actions” (James’ term) such as absent-minded munching, concluding that “these actions seem to roll off in a way that skips intention.” Wegner, \textit{The Illusion of Conscious Will} (Cambridge, Mass.: MIT Press, 2002), p. 130.
his cell; he rattles the bars of his cell because of such a want. Yet he doesn’t rattle the bars in order to get out, for he doesn’t believe for a moment that he can dislodge the bars. His emotion plausibly causes his behavior but not through the mediation of some intention. The question is whether the emotions of fear, anger, and craving typically bypass the will in this way.

I expressed some doubts about this years ago, when the memories of my experiences with an emotionally explosive, near-and-dear relative were still fresh:

“Are any emotions truly free of corresponding judgments that justify them to the agent whose emotions they are? Is any rage truly blind, or any anxiety without its object? Do the emotions that allegedly cause action by ‘short-circuiting’ choice ever proceed except by a chosen letting go, a chosen self-indulgence?”

Time has not mellowed my judgment here.

Separate two considerations, the strength of a motivating emotion, and the suddenness with which it causes the behavior that it motivates. Surely the strength (or intensity) of an emotion, by itself, does not typically result in a by-passing of the will and thus, automatic action. Love for another person can be as deep as you please, commitment to a social cause as passionate as you like, yet when we act on the desires expressive of such emotions our reason is not unhinged, our wills are not bypassed. Which means the temptations to talk of bypassed wills come from the suddenness with which emotion-driven actions can (but need not) be taken up. Such sub-class of emotion-driven actions then joins other forms of spontaneous or impulsive behavior. There are interesting questions to pursue about such behaviors, but they are distant from questions of volitional excuse.

What about the second variation earlier distinguished, the variation where there is no absence of intention (as in the first variation) but where X has two inconsistent intentions to match his two inconsistent desires? The will is not bypassed in such cases, but it is unresolvedly conflicted. Some will deny that such cases are really possible psychologically. Supporting this denial is this kind of thought: because intentions have as their main functional role the resolution of conflict between desires, when no such resolution is reached there can be no intentions

---

52 The example is from Moore, Law and Psychiatry, p. 16.
54 Namely, how do instantaneous reactions to suddenly arising emotions compare with higher order reflexes such as the pain-withdrawal reflex, in terms both of common sense voluntariness and of activated systems in the upper motor pathways above the spine?
formed. Yet I don’t find “conative” dissonance psychologically impossible, at least if one isn’t too fierce in one’s demand that all intentions univocally settle conflicts. After all, cognitive dissonance seem to be a well-accepted phenomenon (where one believes p and yet also believes not-p). Beliefs may not involve practical resolutions of conflicting desires but they do involve theoretical resolution of conflicting bits of evidence, and yet we still credit the claim that cognitive dissonance exists, that is, that there are unresolved conflicts of beliefs. Beliefs in such unresolved conflict each represent a partial resolution, or tendency to resolution; intentions in conflict can be similarly conceived.

More worrisome is how typical (of actions we intuitively regard as compelled) is the psychology here depicted in terms of conflicting intentions. Not very typical is my own guess, given the extraordinary irrationality involved in simultaneously forming contradictory intentions. Further, when such conflicts do exist, do they capture what is distinctively excusing about compulsions? In those rare cases of cognitive dissonance, we do decline to attribute belief to the dissonant agent when his responsibility is on the line.\textsuperscript{55} Similarly, we might also decline to attribute the intention to the agent if needed for intended wrongdoing. But this doesn’t capture the “feel” of what is excusing in cases of compulsion. Nor, as I say, is its occurrence widespread enough to account for but a small percentage of intuitively excused cases.

So how about the third possibility, where we conceive of the compelled actor as one who acts (and intends to act) against what he most wants? This to my ear sounds more congruent with the intuitively excusing nature of compulsions. Unfortunately, however, the proposal is literally incoherent, at least within the conceptual framework that I earlier stipulated. For it violates the truism that what we most want in situations of conflicting desires is what we in fact choose and act on in situations where the alternatives are clear to us. The actor who has belief 2 (about the inconsistency of his desires) and yet who both intends to eat the cake and does just that, cannot most strongly desire at that moment to remain thin.

We could of course go back on my earlier stipulation about strength of desire being measured exclusively by the causal power of desire. One might substitute “degree of intensity of felt cravings,” for example, or “degree of satisfaction upon attainment.” But rather than work with these slippery notions, let me get at what tempts one to use them, in another way.

\textsuperscript{55} Thus, the Model Penal Code ‘2.02(7) recognizes the possibility of cognitive dissonance (where the defendant both believes to a high probability some proposition p and yet believes not-p), but classifies such cases as not intentional or knowing.
Let us start over by modifying the third variation this way. We should keep the desire to remain thin in 1, keep the desire for cake in 1’, and keep the preference order (of cake over thinness) as stated in 1’. Keep the belief (in inconsistency of the two desires) in 2. Keep the intention (to eat the cake) in 3, as well as the act, which is to execute that intention by eating the cake. So far, that is, we have retained all features of the situation that normally lead to a conclusion of responsibility for eating the cake. But now focus on desire 1’. Suppose of it that it is a strong desire as defined earlier. Suppose also that it is an emotion-laden desire, a product of craving, fear, anger, or some other intensely felt emotion. These two features are, as we have seen, insufficient for excuse by themselves. Acting out of a strong and passionate desire to do justice, or a strong and passionate desire to rid oneself of a much-hated enemy, does not affect one’s praiseworthiness/blameworthiness.

Needed for excuse is some third attribute of such desires that make them seem alien to (rather than a part of) the reason-responsive, responsible self. With threats by other people, and with the “threats” of nature involved in cases of natural necessity, the desires to yield come ready-made with a sense of not originating in the self. But with anger or craving distinctive of provocation and addiction, there is no similarly easy external identification of the desire as not being part of the self. Still, even in these latter cases there has been quite a variety of descriptions of what makes such desires not part of the self. Freudians used to call such desires “ego-alien,” meaning that they were experienced as if they belonged to someone else or even to something else, “das es,” or an it, or “id” – in any event, as “not-me.” Early Harry Frankfurt spoke of there being a lack of second order desires, that is, lack of any desire to have such desires as 1’, as well as the presence of a second-order desire to have desires such as the desire to stay thin in 1. Frankfurt raised such (present or absent) second-order desires, not because he thought that they were necessarily stronger than first order desires, but because he thought that persons identified themselves more with such second order desires (and with the desires secondarily desired) than with brute first-order desires. Later Frankfurt sought to capture this greater centrality to self-identity with his notion of wholeheartedness: some desires we

---

58 As Philip Pettit so construes him, in Pettit, “The Capacity to Have Done Otherwise,” p. 25.
wholeheartedly endorse, whereas others (like 1') we do not,\textsuperscript{59} or perhaps we even disvow.\textsuperscript{60} Michael Smith speaks of desires that match the desires the actor \textit{believes} he should have, translated (for Smith) into the strongest desires the actor believe he \textit{would} have if he were fully rational.\textsuperscript{61} Desires like 1' are not, for Smith, the objects of such evaluatively hypothetical beliefs. They may even be acknowledged by the agent to be defective desires in that they conflict with the desires he believes he would have in greater strength if he were fully rational. Victor Tadros more recently speaks of desires that are not “accepted [by the agent] in light of the agent’s values.”\textsuperscript{62}

Painting with a somewhat broad brush, I see these various formulations all referring to roughly the same thing: some desires are tightly woven into an agent’s view of himself, and others are not. When a strong, emotion-laden, not-identified-with-self desire conflicts with a less strong, probably less emotion-laden, but more identified-with-self desire, and wins (in the sense of determining intentions and actions in conformity with it), the possibility of excuse for that behavior arises.

I am certain that the psychology depicted by Freud/Frankfurt/Fingarette/Smith/Tadros and many others could be fine-tuned in various ways; but however that is done, it seems to me to be both plausible as a psychology and co-extensive with a clump of the cases we are tempted to regard as cases of volitional excuse. Which is not (yet) to say that such conditions are in fact excusing.

One of my old worries about the excusing nature of these ego-alien desires I still fret about, more strongly in some cases of alleged compulsion (such as many provocation and addiction cases) than in others (such as many cases of duress and necessity).\textsuperscript{63} This is the worry that we as moral agents have limited normative power to map out the domain of excuse for ourselves by our self-identifications. Freud once scornfully remarked that “the physician will leave it to the jurist to construct for social purposes a responsibility that is artificially limited to the metaphychological ego,” and that this would be to “disregard the evil in the id” and “not

\textsuperscript{60} On avowal and disavowal of desires, and emotions, see Herbert Fingarette, \textit{Self-Deception} (London: Routledge, 1969).
\textsuperscript{61} Smith, “Responsibility and Self-Control.”
\textsuperscript{62} Tadros, \textit{Criminal Responsibility}, p. 343.
\textsuperscript{63} Moore, \textit{Placing Blame}, p. 556.
make my ego responsible for it.”64 Our “ego-alien,” not-second-order desired, negatively evaluated desires are after all, our desires, as are the emotions that may produce them. That our own self-identification excluding such desires and emotions from self can make us excused from responsibility for the actions they produce, is troubling. Seemingly the size and boundaries of our moral agency is not up to us in the way or to the degree that this doctrine of excuse suggests.

Decades ago I tentatively voiced this worry about this account of volitional excuse. My current take on the worry is that it doesn’t eliminate volitional excuse, although it helps to cabin its application. In those familiar cases of provocation that populate American criminal law casebooks – where with some frequency pathetic males kill their female victims out of anger, disappointment, feelings of rejection, sexual frustration, and the like65 – I have no sense of excuse. This, despite strong, emotion-driven desires with which the agent does not self-identify, motivating such killings. For the identification of self by such agents in a manner that excludes (negatively evaluates, etc.) such emotions and desires, is too self-forgiving to credit. It is right up there with the child’s, “the devil made me do it,” in terms of legitimate excuse. Still, perhaps what prevents judgment of excuse in such cases is the lack of genuineness and sincerity in the narrow identifications of self by such actors. Perhaps what is really true of such unexcused individuals is that their actual sense of self is broader than they claim when apprehended, in that they identify with their emotions of injury, insult, frustration, anger, disappointment. They feel entitled to feel in such ways and thus don’t actually disavow such feelings, as much as it behooves them at trial to pretend otherwise. Truly unbidden and surprisingly alien emotion/desire combinations may indeed be excusing, despite the plethora of false imitations that show up in court.

A fourth and final model of volitional incapacity operates through belief as well as through desire. In some cases one’s inability to will what one knows he should will is made possible by the degrading of the belief that the two desires are inconsistent. Sometimes, that is, the tempting desire causes its subject to believe that he can “have his cake and eat it too,” even though generally he knows that that isn’t true. This is the familiar phenomenon of wishful thinking, where a belief in some proposition p arises because the subject so badly wishes p to be true. If one is tempted very much to eat the cake in front of one now, it may be that one’s belief

65 E.g., People v. Casassa, 49 N.Y. 2d 668, 404 N.E. 2d 1310 (1980).
about the inconsistency of remaining thin and eating such desserts (and one’s belief in the need for lines to be drawn to prevent small breaches of discipline to become major deviations) get suspended. A familiar enough form of yielding to temptation is one whereby one believes at the moment of acting that there is no significant cost to the maintenance of one’s long-term, controlling goal of remaining thin, by satisfying one’s tempting desire now.

Such wishful thinking, like self-deception generally, has the flavor of active manipulation of one’s own beliefs. As such it is not clear how incapacitating it is, and thus unclear how excusing it is. After all, one is in such cases believing what one wants to believe because one wants to believe it. Yet the paradox of self-deception is that one is also a deceived victim of deception as much as agent of deception. Although there may well be therapeutic efficacy achieved by taking responsibility for such maneuvers, in truth we are not morally blamable for actions done under mistaken beliefs even when such mistakes arise in this way.66 The upshot is that yielding to temptations via the mechanisms of wishful thinking may also provide one with some moral excuse for such wrongful actions.

Notice that so far we have said nothing about how strong, emotion-laden, ego-alien desires incapacitate actors from making better decisions. Once we left the bypassing of will story of the first version of this excuse, we left the only story we had about incapacitation. Even if strong and emotion-laden desires can and should be identified as alien to my moral agency in the ways indicated, how do such ego-alien emotion/desire sets incapacitate what is clearly “me” (no matter how narrow may be the boundaries of self), i.e., my deciding self (or “ego”). We have not yet approached the line distinguishing the agent who could have but did not resist tempting desire, from the agent who couldn’t have done other than he did in yielding to such desire. That issue will occupy us in a succeeding section. But first we need to deal with the other major possibility for volitional excuse, viz, where one forms the right intention but is unable to act on it.

D. Inability to Execute the Right Intention

The psychology for this second major kind of possibility has both desires 1 (to be slim) and 1’ (to eat the cake) in it; also the belief 2 as to the conflict between these desires; also the preference order 1” where desire 1 (to remain slim) is stronger than desire 1’ (to eat the cake); and

the intention 3 (not to eat the cake). If the actor in fact refrained from eating the cake as he intended, then this would be a normal case of acting on conflict of desires with no hint of excuse arising. But suppose X eats the cake, contrary to his intention. Then X is classically considered as an akratic, that is, one weak of will.

Many would deny that anyone actually behaves this way. Yet again, we should separate criticizable irrationality from psychological impossibility. The akratic is indeed highly irrational; but that does not mean there are no such cases. One ground for thinking such cases to be impossible and not merely irrational, is the truism discussed before, that we necessarily act on the strongest desire. Yet such necessity arises out of a definition of stronger desire behaviorally, as that desire in fact acted upon in situations of known conflict with other desires. It wouldn’t take much alteration to make akrasia possible: define strongest desire as the one the actor in fact chooses in situations of known conflict with other desires. The akratic’s choices (intentions) are in line with this modified truism. For his choices are fine; it is only his behavior that is disappointing.

Another temptation is to think that an intention that produces the opposite of the action intended – when nothing intervenes to induce a change of mind, and no mistakes in beliefs about means of execution are present – goes against what intentions are. On a dispositional view of intention, it is part of the “logic of the concept” (as the ordinary language philosophers used to say) that one does the act intended when the occasion to do so arrives and nothing relevant has changed in the actor’s mental states. Yet this old view of intentions is too behavioral to be credible; “intention” refers to a natural kind whose deeper nature is functional and physical, not phenomenological and not behavioral. Van Gogh can intend to be a great artist, and this can cause him to cut off one of his ears. This is wildly irrational, but not psychologically impossible.

A third worry about this picture of weakness of will is that such weakness occurs only in automatic actions. This, because of the absence of any intention mediating between desire and action. True, in such cases there is an intention such as 3, an intention to refrain from eating the cake; and true, that intention can serve as a causal intermediary between the strongest desire in 1 (to remain slim) and the action (of eating the cake) despite the mismatch of act done both to act most desired and act intended. Yet won’t most such cases where this array of mental states is present be cases that tempt us to say that the weaker desire 1 (to eat cake) directly caused the
action desired? And if this is true, such actions will be part of the “automaticity of everyday life” earlier discussed.

The worry is that many cases of plausibly compelled and even obsessive behavior are not sudden yieldings to temptation. As J.L. Austin observed, one can take the second dessert at High Table (which one knows one shouldn’t have) with delicacy, deliberation, and graceful slowness; yielding to temptation need not always be, and often is not, the wolfing down of such dessert.67 The latter is the exaggerated depiction of yielding that is the stuff of grade B movie scripts.

None of these three problems to my mind rule out the possibility of weakness of the will as I have depicted it, where the strongest desire to do the right thing is realized in a choice (intention) to do that very thing, and yet (with no other mental state intervening), the actor does the opposite. Even though deeply irrational such behavior is psychologically possible. Yet surely such deeply irrational behavior is comparatively rare, indeed, too rare to capture the range of cases we intuitively think of as weakness of will.

There is a better conceptualization of the phenomenon, one that can be seen by attending to these examples of Thomas Schelling (a noted game theorist and economist looking for a rational consumer whose preferences would be worth maximizing in a utilitarian calculus):

“How should we conceptualize this rational consumer whom all of us know and who some of us are, who in self disgust grinds his cigarettes down the disposal swearing this time he means never again to risk orphaning his children with lung cancer and is on the street three hours later looking for a store that is still open to buy some cigarettes; who eats a high calorie lunch knowing that he will regret it, does regret it, cannot understand how he lost control, resolves to compensate with a low calorie dinner, eats a high calorie dinner knowing he will regret it, and does regret it; who sits glued to the TV knowing that again tomorrow he’ll wake early in a cold sweat unprepared for that morning meeting on which so much of his career depends; who spoils a trip to Disneyland by losing his temper when his children do what he knew they were going to do when he resolved not to lose his temper when they did it?”68

67 Austin, “A Plea for Excuses.”
What Schelling’s familiar examples from daily life suggest to me is not actions that go against both strongest desire and intention, as modeled above. Nor is it what seems to tempt Schelling himself (which is the simultaneous dual intention model I examined and put aside earlier). Rather, Schelling’s examples suggest that we go diachronic:69 keep the truism that we necessarily do that which we most strongly desire, keep the match between object of intention and object of strongest desire, and keep the match between action done and action intended. But see Schelling’s agents as oscillating over time between sets of mental states and actions, each of which obey these requirements. So at \( t_1 \), the smoker:

1. Desires (not to risk lung cancer by smoking)
2. Believes that it is not the case that he can both smoke and not risk lung cancer

Therefore: 
3. Intends (not to smoke)
4. Does not smoke

Yet at \( t_2 \), the mental constellation of mental states change, resulting in the opposite action. The smoker still has desires 1 and 1' and belief 2; but he:

1’. More strongly desires to smoke than he desires not to risk cancer

Therefore: 
3’. Intends to smoke
4’. Does smoke

And then at \( t_3 \), the period of immediate regret, he oscillates back to the first constellation of mental states.

Notice that none of the three objections earlier raised apply to this diachronic conceptualization of weakness of will. There is no need to amend the truism about strongest desire necessarily being the one acted upon. There is no need to qualify the view that ties


“[W]e cannot identify weakness of will simply by looking to see whether at each moment the agent’s acts correspond to his deliberative conclusions at that moment; we need to survey several aspects of the agent’s history overtime, including…the frequency and reasons for ‘changes of mind.’”

dispositions to behave to intentions (for the agent is disposed to behave in accordance with the objects of his intentions at each time). There is no need for the act of smoking to be sudden or automatic, because it is not directly caused by desire but is rather guided by an appropriate intention.

Another objection does apply, however. This agent’s intentions – both 3 at t₁, and 3’ at t₂ – are decidedly non-sticky. Unlike ordinary, sticky intentions, non-sticky intentions do not preclude constant re-evaluation of what the agent most wants to do (or thinks, all things considered, that he should do). Such non-stickiness is criticizably irrational. But it is not only psychologically possible, but surely it is just as common as Schelling plainly thinks it is.

Moreover, isn’t this a good match to the idea of a will that is weak on a given occasion? The decisions (choices, intentions) of such a weak-willed person don’t control his behavior much into the future because they themselves are so constantly subject to being changed. Such lack of psychological commitment to the non-reconsideration that having an intention rationally commits us to, well unpacks the idea of a will that is weak.

Suppose that this is a plausible psychological picture of weakness of will. Is such weakness – when exhibited by an agent on a particular occasion of intentional wrongdoing—excusing? As a first cut, surely the intuitive answer is no. Such weakness is not only a defect of rationality, it is also a moral shortcoming. When St. Paul complains in Romans vii: that “the good which I want to do, I fail to do” and that “what I do is the wrong which is against my will…,” he was not exonerating himself. Such weakness to do what one knows is right has perhaps a contemptible cast to it that fully affirmed and willed evil does not, but both on their face are morally condemnable, not excusing. True, we have limited capacities to strengthen our will, either in general (will-power exercises to build our resolve?) or in particular cases (willing ourself to be stronger of will?). But that is true of our ability to shape our desires too – yet no one thinks that my insufficient concern for others, my hatred of some virtuous person, or my fondness for watching others suffer, excuses me just because these attitudes, desires, or emotions are difficult to eliminate or even substantially change very much. Some aspects of who

we are grounds our blameworthiness for our actions even when those aspects are not subject to our willing them to be otherwise.

So as a first cut a rejection of there being any general excuse of lack of will-power seems appropriate. I take it that Anglo-American criminal law recognizes this moral truth in its doctrines of duress and provocation. The Model Penal Code allows the excuse of duress only when the threats are such that a “person of reasonable firmness” would have been unable to resist them. Such a restriction eliminates weakness of will as a legal excuse (although it leaves intact the ego-alien, emotion-driven, strong desire model of compulsion earlier discussed). Similarly, the common law’s partial defense of provocation requires that the provoking act of the victim be such as would make a “reasonable man” lose his powers of self-control over his anger. One of the attributes that makes a person “reasonable” in this context is that he has the power to control emotions (like anger) possessed by a person of reasonable firmness. The hot-tempered, impulsive, pugnacious, emotionally explosive, unthinking brutes get no excuse under such a standard, no matter how deeply and how demonstrably they lack the power to control their emotions because their will to do so is weak.

Yet apart from criminal law’s confirming morality’s denial that there is any general excuse of weakness of will, the criminal law more interestingly evidences a more subtle moral truth: for sometimes weakness of will – inability to effectuate one’s intentions formed at an earlier time – is an excuse. For sometimes the lack of will power is not a moral defect in the person who lacks it. An easy example is intoxication. It is common for intoxication to loosen the inhibitions of the intoxicated person. In such a state he has less control over his emotions of fear or of anger, and of the desires that they spawn, with the result that he can maintain his resolve (earlier intention) less successfully over time. When the intoxicated state is not his fault – as it is not in involuntary intoxication – then he has a more plausible, perhaps partial excuse of weakness of will.

Youth is another easy example. The time-discounting is steep, the impulse control poor, for young people as opposed to adults. And this is not their fault – being young is not a moral defect (however undesireable/unattractive it may sometimes be, particularly in American teenagers). The chronologically immature have not yet had a fair opportunity to develop into adults

---

73 Model Penal Code '2.09.
74 Duff, “The Virtues and Vices of Virtue Jurisprudence.”
of whom we may fairly expect a higher standard of self-control. So young wrongdoers too have some excuse of weakness of will.

It is not that the involuntarily intoxicated, the young, and others with blamelessly lowered abilities to maintain their resolve in the face of fear, anger, cravings, or other emotional states, get a complete pass. For they only have a lesser capacity to control themselves, not no capacity. Anglo-American criminal law recognizes this last fact by asking whether the young, the intoxicated, etc., did as well as could fairly be expected of one with the lessened capacity typical of those similarly young or drunk. But where there is no unexercised capacity (as judged by this lesser standard), then there is excuse.

So there is some room for a viable excuse of weakness of will. How much depends on how many conditions there are where two things are true: (1) the power of self-control is lessened from what we normally demand of persons generally; and (2) it is not a moral defect in such persons to have such lessened powers of self-control. From the English case law: Impotency? Gender? Combat training? Past racial discrimination? Past domestic abuse? Mental disorders of various sorts? Etc.

E. Summary of the Folk Psychological Models of Compelled Choice and Action

So we have arrived at two psychologically plausible (and sufficiently common) models of when someone is compelled. The first locates compulsions in strong, emotion-driven, ego-alien desire; the second locates compulsion in the weakness of the executive faculty that is supposed to mesh such desires in with all else that an agent desires, the will. A compelled person is one who is either unable to form an intention to act on the desire with which he identifies himself, or is unable to execute that intention (despite being able to form it) because through no fault of his own his will is unable to preclude itself from constant reconsideration.

These two are the main variations. Yet notice that there are four other models having some range of plausible application as well. That means there are a total of six models of excuse here. Four of these models focus on cases of inability to form the intention to do the right thing; two focus on cases of inability to execute such an intention. One way to conveniently summarize

---

75 The special standard accorded various classes of wrongdoers has received much attention from criminal law scholars. See, e.g., Tadros, Criminal Responsibility, pp. 349-358.
76 Culled from Tadros, Criminal Responsibility, pp. 349-358.
the first five models (the synchronic ones) is by contrasting their folk-psychological attributes with those of the moral case of responsible action. Thusly:
### MODELS OF INCAPACITY TO FORM OR ACT ON THE RIGHT INTENTION

<table>
<thead>
<tr>
<th>Normal Case of Responsible Action</th>
<th>Model-1 Bypassing of intention</th>
<th>Model-2 Conflict of intention</th>
<th>Model-3 Ego-alien desire</th>
<th>Model-4 Wishful thinking</th>
<th>Model-5 Synchronic weakness of will</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Long range, controlling desire (e.g., to remain thin)</td>
<td>1. Same</td>
<td>1. Same</td>
<td>1. Same</td>
<td>1. Same</td>
<td>1. Same</td>
</tr>
<tr>
<td>1’. Short range, immediate, tempting desire (e.g., to eat cake now)</td>
<td>1’. Same</td>
<td>1’. Same</td>
<td>1’. Same except the tempting desire is: a. Strong b. Emotion-laden c. Ego-alien</td>
<td>1’. Same</td>
<td>1’. Same</td>
</tr>
<tr>
<td>1.” Preference order: tempting desire (1’) stronger than controlling desire (1)</td>
<td>1.” Same</td>
<td>1.” Same</td>
<td>1.” Same</td>
<td>1.” Preference order: tempting desire (1’) may be stronger or weaker than controlling desire (1)</td>
<td>1.” Tempting desire (1’) is weaker than controlling desire (1)</td>
</tr>
<tr>
<td>2. Belief in inconsistency of satisfying both controlling desire (1) and tempting desire (1’)</td>
<td>2. Same</td>
<td>2. Same</td>
<td>2. Same</td>
<td>2. Belief in consistency satisfying both controlling desire (1) and tempting desire (1’)</td>
<td>2. Same</td>
</tr>
<tr>
<td>3. Intention in line with strongest desire (e.g., to eat cake)</td>
<td>3. Absence of any intention in line with strongest desire</td>
<td>3. Same; but plus: 3’. Intention not in line with strongest desire (e.g., intention not to eat cake)</td>
<td>3. Same</td>
<td>3. Intention in line with tempting desire (1’)</td>
<td>3. Same (e.g., actor intends not to eat cake)</td>
</tr>
<tr>
<td>4. Action done in conformity with intention (e.g., actor eats cake)</td>
<td>4. Same</td>
<td>4. Same</td>
<td>4. Same</td>
<td>4. Same</td>
<td>4. Action is not done in conformity with intention (e.g., actor eats cake)</td>
</tr>
</tbody>
</table>
The last model, that of diachronic weakness of will, is simply a temporal vacillation between the mental states in model-5 (when the actor most wants and intends not to eat cake) and the mental states in the normal case of responsible action (where the actor most wants and intends to eat cake), where the latter mental states govern the act done (the eating of the cake).

Notice that settling on plausible psychological models of compulsion does not yet draw the crucial can’t/won’t distinction. I just formulated the two kinds of models in terms of agents who are unable to form the needed intentions or who are unable to act on such intentions if formed. What we haven’t done is unpack these inabilities; in particular, we haven’t yet drawn a line distinguishing those unable to form or act on certain intentions, from those who simply do not form or act on such intentions. For this we need to turn to some analysis of capacities that allows us to draw the needed distinction.

III. FRAMING THE APPROPRIATE COUNTERFACTUAL CONDITIONALS THAT ANALYZE THE TWO MODELS OF INCAPACITY

A. The Relevant Conditions for Choosing and Acting Otherwise

With the excursion into the folk psychology of coercion complete, I turn to the question of how we are to separate cases of morally excusing incapacity – where one could not have done otherwise – from cases of non-excused bad behavior – where one could have done otherwise but simply wouldn’t do so.

The key to the distinction lies in the counterfactual analysis of ability. While not universally accepted, it is a plausible analysis of ability to analyze, “X could have A-ed” in terms of the counterfactual, “X would have A-ed if C,” where “C” represents a change from the actual world (in which X did not A.) Such changed worlds philosophers since Leibniz have called, “possible worlds.” To say of a track star who in fact lost a given race, “he could have won that race,” might for instance mean, “in the possible world where he tried much harder than he did in the actual world he would have won.”

Elsewhere I have defended the counterfactual analysis of ability against the roughly ten objectives that have been advanced against the analysis in philosophy these past fifty years. Let

---

79 Moore, “Compatibilism(s) for Neuroscientists;” Moore, “Stephen Morse and the Fundamental Psycho-Legal Error.”
me just mention one such objection, perhaps the most famous, this one advanced by J.L. Austin. Austin’s well-known example was that of the missed golf putt. Of a skilled golfer who misses any easy putt, we might truthfully say, “he could have made that putt.” Yet, Austin queried, if this “could” is translated to, “the golfer would have made that putt” that conflicts with the blunt fact that he didn’t make it. Therefore, Austin concluded, we cannot translate “he could have made it,” to “he would have made it.”

It is easy to miss Austin’s point if we fail to attend to the distinction introduced earlier, that between general capacities and particular capacities. Suppose we truthfully say of a skilled golfer, “he can make putts” (within certain degrees of difficulty). If those who would reduce capacities to counterfactuals are accused of reducing this to, “he would always make putts (within this degree of difficulty),” then Austin is plainly right; even the most skilled golfers (who “can” make such putts) don’t make them all. But equally plainly, the reductionist can amend his reduction base from, “would make such putts all the time,” to, “would make such putts most of the time.” But this imagined exchange is all by-the-by to the point of interest to both reductionists and Austin, which was a point about capacities on a particular occasion.

Suppose what we want to know about the golfer is not whether generally he can make putts of this degree of difficulty – we are conceding, let us suppose, that he is a generally skillful golfer and that in general he can make such putts. What we rather want to know is whether this golfer, at the time at which he missed a certain putt, could have made that putt on that occasion. In asking this latter question what we want to know whether it was possible for that golfer to have made that putt; that such a putt was within the degree of difficulty of putts he generally has the ability to make, is relevant but not determinative of the particular question.

A reductionist reduces the “could he have made that putt” question into, “would he have made that putt if C?” Austin’s objection (that the golfer who truly could have made that putt nonetheless missed that putt in the actual world) now seems to miss the point – of course, one might say, the golfer missed the putt in the actual world, for something, call it “not-C,” prevented him from making it! But with not-C removed (i.e., with C present), he would have made it, is a sensible thing to think. Latter day Austinians with whom I am familiar tend to deny that there were some identifiable conditions (not-C) that prevented the successful putt on this occasion.

“The golfer just missed,” they often say; or the golfer missed “for no reason,” as Austin himself put it. But isn’t this pretty bogus? In a deterministic world there is always a reason why a skilled golfer misses an easy putt. The ground tilted in a way not visible to the putter; the grass was of uneven cut; the ball was off-center, of uneven density, or was not perfectly round; he obliquely hit the ball so that it spun, something he rarely does; etc. Call whatever explains the miss, “not-C.” Now conditionalize on the absence of not-C, i.e., on C: “He could have made the putt” now is translated to, “He would have made the putt if C.” And that looks pretty good, does it not? Every ability has success conditions; this golfer’s ability does too, which in the context of this particular putt were C. In conditions C, he would have succeeded.

Notice that all that has been established so far is that capacity statements are disguised counterfactuals of some form. That does not establish that the specific counterfactual Moore proposed for volitional excuse – the actor could not have done otherwise even if he chose (or tried) to do otherwise” – is the right one. After all, capacity statements generally are highly context-dependent on the conditionals appropriate on different occasions. The statement about the track star who “could have won” the race that he in fact lost, could involve any number of conditionals, depending on the circumstances. Perhaps: “if he had lengthened his stride,” or, “if he had not been so distracted;” etc. Indeed, any state of affairs that in fact prevented him from winning could be used to form a counterfactual, appropriate to some occasion of utterance. The Moore-like conditional – he would have won if he had tried harder (or chosen to) – is only one amongst myriad possibilities.

Our present context is that of assessing the responsibility of actors who either acted against their own best judgment (but in the name of their strongest desire) or acted against their strongest desire and its resulting intention. Let us focus on these two dominant folk-psychological models of compulsion that we have just explored. These models are the place to start in deciding upon the appropriate conditionals.

On the first of these two dominant models of compulsion – the craving model where ego-alien desire produces a choice to eat cake (and an eating of cake) – we first need to settle on the consequent of the relevant counterfactual(s). Do we want to ask, “could he have not acted so as to eat the cake?” Or: “could he have chosen not to eat the cake?” For this model of compulsion, where the problem lies between desire and choice rather than between choice and act, the latter

---

81 Id.
question is our primary concern. “Could he have chosen otherwise?” is thus our relevant question. The answer to the question, “could he have acted otherwise,” is not without interest for us; but that answer will follow the answer to the question about choice in the cases we are here isolating with this model of compulsion (where choice causes the act chosen to be done).

In fleshing out “C,” the minimal condition of what must change (as we leave the actual world for the possible world in which we are to see if the actor does or does not do action that he did do in the actual world) certainly includes a heightened desire to resist the temptation. This is the central idea behind the relevant counterfactual, “the actor would have refrained from doing A if he had wanted to.”

This initial specification of C requires further specification in two dimensions. One is a matter of quantity: how much (more) does the actor want to resist the temptation in the possible world we are examining? A traditional answer is an extreme one here: only actors who still do A despite facing the strongest possible incentives not to do A (either in terms of rewards foregone or punishments visited upon them for doing A), are said on this view to be unable to refrain from doing A. In fleshing out C in this traditional way, then, A’s controlling desire not to do A must be as strong as any desire he has. (I shall soften this traditional answer later.)

This raises the question of how such an enormously powerful desire could ever lose out in determining what the actor chooses to do. Here is where the four models of volitional incapacity come in. If no overall balancing of competing desires is possible – either because no intention is formed (the first, bypassing model), or because the only intentions that are formed are those that replicate the conflict in desires (the second, conflict model), or because the alien nature of the tempting desire refuses integration enough to be balanced off against the controlling desire – then the tempting desire might cause behavior irrespective of the heightened strength of the controlling desire. Analogously, where an actor believes he can “have his cake and eat it too,” the greater strength of the controlling desire will be without effect in determining his choice – for he sees no inconsistency in satisfying his tempting desire while also satisfying the demands of his controlling desire.  

---

82 We are now in a position to stipulate more completely the meanings to be assigned to strength of desire, and strongest desire. The strength of a desire is its tendency to win (in the determination of choice and behavior) vis-à-vis other desires; the strongest of a pair of desires is the one which would win in situations of normal conflict, i.e., situations that are not models 1-4.
These are the four ways in which the controlling desire is denied purchase in the determination of choice, the common upshot being that that desire’s greater strength is without effect. In such possible worlds the actor might still have chosen to eat the cake even if he desired very much more to remain thin, in which case we say, he couldn’t have chosen otherwise even if he wanted to do so.

This sounds disarmingly simple. Which it is, but not that simple. To see why will introduce the second dimension of further specification needed. As David Lewis showed pervasively, when we change the world from its actual state (not-C) to some other, possible state C, some other changes (beyond the change from not-C to C) must be contemplated. If the controlling desire is stronger in the possible world than it was in the actual world, something else has to differ between the two worlds. Even if one steadfastly refuses any change in any other event or state prior to C, so that exactly the antecedents that produced a weak desire in the actual world, produced a stronger desire in this possible world, something will have to have changed, viz, the scientific law(s) connecting those antecedents to C. This is what Lewis called “a minor miracle,” because in such cases some scientific law inexplicably must not have held over some time interval prior to C. Alternatively, perhaps the laws were intact but some of the antecedents of the weak desire changed so as to produce the stronger desire. There are many discrete properties that give a desire its causal power on a given occasion. Vividity of awareness of the desire; degree of experienced cravingness for its object; remembrance of the degree of satisfaction felt for similar desires when satisfied; general mood; dominant image of self at the moment (slim, fat, etc.); accidents of association with other desires; social reinforcement of the desire by friends or others; the aphrodisiac effect of other desires felt at that moment; etc., can all affect how effective a desire is in realizing itself through choice. Perhaps one or more of these discrete attributes of strength of desire also changed.

Lewis held that in judging how close a possible world is to the actual world, we should regard minor miracles and local changes of state as making for greater closeness over and against revisions of many scientific laws or changes in state of whole space/time regions.

---

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.

To summarize: the counterfactual about X not choosing to eat the cake is tested in possible worlds close to the actual world, save that in those worlds: X desires to remain thin a great deal more than he does in this world; and whatever else has to change to generate such strengthened desire by X has also changed. When X would have chosen not to eat the cake in such close possible worlds, then X could have chosen not to eat the cake in this, the actual world; when X would have chosen to eat the cake despite a greatly heightened desire to remain thin (but because of the disabling of such strong desire by one of the four mechanisms above specified), then X could not have chosen not to eat the cake in this, the actual world.

Turning to the second model of compulsion, that of weakness of will, we also here need to first settle on the content of the consequent clause of the relevant counterfactual. For this model of compulsion (unlike the first set of models), we want to focus on how the actor would have acted, not (in the first instance) on what he would have chosen. For the problem in this model of compulsion lies between choice and action, not between desire and choice as in the first set of models. Accordingly we want to ask whether actors like X would have acted otherwise, not whether they would have chosen otherwise – for with the cases here considered, their choices (not to eat cake, e.g.) were just fine.

It is thus Moore’s original kind of conditional on which we should thus here focus: “Would X have not eaten cake if C?” Moore’s analysis of the “C” was “Would X not have eaten the cake if he had chosen not to do so?” In cases where X does eat the cake despite a choice (or intention or a trying) not to do so, X could not have done otherwise, on this Moorean construal of this phrase. Moore’s counterfactual then accurately separates those who couldn’t do otherwise from those who simply wouldn’t, as Moore intended it to do. In both categories of weakness of will – the synchronic and the diachronic models – the actor’s choice does not cause the action
chosen, and he thus suffers from a volitional incapacity. Lacking is any causal impact (the synchronic case), or at least any such impact sustained over time (the diachronic case).

The problem is that this form of volitional incapacity does not accurately capture situations where the actor is morally excused for such incapacity. Moore’s counterfactual is too broad to serve that purpose. The truly excusing cases are only a subset of the cases captured by the Moorean counterfactual.

We thus need to supplement Moore’s counterfactual with the stipulation that an actor is excused only if (1) he could not have done otherwise, in Moore’s sense; and (2) his incapacity in this regard is not a moral defect in his character. While an actor may be incapacitated by weakness of will, he will only be excused for actions he could not help doing if his inability in this regard is not his fault.

One might be tempted to capture this second feature of excuse with a second order capacity, such as:

“X is at fault for his weakness of will if and only if at some earlier time he could have strengthened his will.”

The right hand side of this biconditional in turn might be given a counterfactual interpretation, so that X could have strengthened his will just in case:

“X would have had a stronger will if he had _______” (where the blank is to be filled in with indirect strategies for strengthening will or for not weakening it by depletion of will-power resources). 85

Yet this is by-and-large a bootless enterprise. Lack of will power can be a moral defect in the person who has it, even though there was nothing he did to bring it about and nothing he could have done or omitted to do to strengthen it. That his intentions have no “stickiness” — no

preclusive effect for future decisions and the motivations behind them – can be a moral defect in him without his doing anything to make himself that way. It is a bit of “constitutive luck” built into moral blameworthiness that incapacities on a given occasion do not excuse if they are an “unlucky” defect in this sense.

Rather than layering a second counterfactual on top of Moore’s, it would be preferable to modify Moore’s counterfactual by altering Moore’s specification of “C.” Why not say that X could not have done other than A just in case:

X would not have done A if he had chosen not to do A, which choice (intention, trying) had the minimal stickiness required by morality for a person of X’s type. If X still would have done A in such a possible world, despite his making a choice not to do A (that was as sticky as it needed to be so as not to evince a moral defect in one with X’s salient characteristics (youth, etc.)); then X could not have done other than A in this, the actual world. Whereas if X would not have done A in the possible world where his choice not to do A had the requisite stickiness, then X could have done A in this, the actual world.

By so specifying further Moore’s counterfactual, we align the can’t/won’t line here too with the line of excused/responsible. To be sure, it is a bit of stipulative redefinition – for in a perfectly ordinary (and even compatibilist) sense, some unexcused actors couldn’t have done other than they did. Yet our aim has been to tame compatibilism, by keeping its analysis of incapacity within morally plausible dimensions, and my stipulative redefinition does this.

B. Disquieting Implications of the Possible Worlds Analysis of the Counterfactuals Involved in Capacities

I suspect that the foregoing analysis of the can’t/won’t distinction will generate some disquiet among some readers. There are two sources of such disquiet related to the possible worlds interpretation of the counterfactuals involved in the two kinds of folk psychological models of compulsion. One stems from the fact that the possible worlds analysis is incapable of unpacking a binary distinction between what one can’t do versus what one doesn’t do; rather; the analysis replaces any such binary distinction with a scalar distinction, one whereby ability to do otherwise is a more-or-less affair, a matter admitting of degrees.

In truth we talk in law as well as in ordinary speech as if an accused either did or did not have the capacity to have acted other than he did; the choice between whether he couldn’t have done better versus he simply wouldn’t do better, is treated as binary. One sees this in
conservatives regularly producing extreme cases, where the controlling desire in the first model is about a life or death matter, or the intention not to do some wrong in the second model is reinforced by considerations equally draconian. E.g., Kant:

“Suppose that someone says his lust is irresistible when the desired object and opportunity are present. Ask him whether he would not control his passion if, in front of the house where he had this opportunity, a gallows were erected on which he would be hanged immediately after gratifying his lust. We do not have to guess very long what his answer would be.”

Like Stephen Morse’s earlier “gun at the head” or the proverbial “policeman-at-your-elbow” hypotheticals, Kant assumes that one positive example of a person’s overcoming some compulsion suffices to show capacity to have chosen or acted otherwise. The assumption is that capacities are such that one who truly lacks capacity cannot control himself no matter what the incentives, and that all others have the capacity to control themselves without distinction between them.

If the possible worlds analysis of the counterfactuals involved with capacity is correct, capacity/incapacity is a matter of degree, and rarely if ever will someone completely lack capacity to refrain from his own intentional actions and the choices that precede them. For notice that degree vagueness in capacity is inevitable in light of the degree vagueness inherent in finding possible worlds “close” or “not-close” to the actual world. The closer the possible worlds in which someone would have done otherwise, the more ability he had to do otherwise in this, the actual world; the more remote the possible worlds in which someone still would not have acted other than they did, the less ability he had to do otherwise in this, the actual world.

Moreover, the fineness of grain in differences (in strength of controlling desire or in stickiness of intention) that make a difference in behavior guarantee a very smooth, finely individuated continuum along which one’s abilities in this regard are measured. There is no lumpy function here. We can only say that some have more capacity to have done otherwise on a given occasion, and some have less. And it takes some seemingly arbitrary stipulation to say where on that matter of continuous variation “can’t” begins and “won’t” ends.

Yet isn’t this implication of the possible worlds analysis – the scalarity of capacity – a virtue, not a vice, of the analysis? On reflection, surely we all think that capacity to choose and to act other than we did is a more-or-less affair, both between people and also between the choices of a given person at different times and within different choice sets. As a legal matter the criminal law may impose a binary categorization on this by-degree continuum, classifying all cases as either “can’t” (excused) or “won’t” (responsible). 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. 87

It is true that if one imposes such bivalent legal remedies on what in nature is a matter of continuous variation, one faces the nagging problem of stipulating an embarrassingly precise point to which the change of remedy is to be attached. One of the hallmarks of a conservative mind set when confronted with this problem is to draw the needed line in some clear place, even if it clearly is not the best place. Consider the Aristotle statement with which we began: the “can’t” of volitional excuse is reached, Aristotle said, only in situations where “no one could withstand the pressure.” No amount of increasing the strength of the controlling desire could change the choice of such an actor, in other words; and only these extremely coerced choices are excused. In modern times in insanity contexts the same extreme standard is recommended: if the accused would have done what he did “with a policeman at his elbow,” then but only then did he lack the capacity to have done otherwise.

These are clear lines, but they draw a harsh border. One could imagine a morality this harsh, fitting perhaps for a race of beings made of sterner stuff than are we. Surely no one really subscribes to this ultra-conservative drawing of the line, however. A line less clear in its delineation, but more sensible in its allowance of excuse, is more plausible. 88 If 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 he did, then the level of excuse is reached. Wishy-washy and vague, for sure, but it at least this is in the vicinity of the line a decently compassionate


88 See Stephen Morse, “Against Control Tests,” p. 453: It would be “too demanding” to require aversive consequences of the “gun at the head” kind; for “if the agent can control himself in such circumstances, it would not follow that the agent could control himself in ordinary circumstances.”
morality draws. I take it that when the criminal law excuses insanity when it “substantially” impairs behavioral controls,89 or duress when the seriousness of threat passes the point tolerable to a person of “reasonable firmness,”90 or provocation when both the level of anger and the behavioral response to it are “reasonable” in light of the situation provoking it,91 the law prefers wishy-washy but correct to clear but incorrect.

The second source of disquiet with the possible worlds analysis of the counterfactuals involved with capacity lies in the nature of the possible worlds analysis itself, not with its implications for capacity. This worry has several strands. One is ontological: do we have to subscribe to the “lush ontology”92 of David Lewis’ modal realism93 to make sense of measuring degrees of closeness of possible worlds? If so, this would be troublesome, given the extravagance of Lewis’ ontology of possibilia. My own view is that we can avoid the extravagance. Even those of us who are older, Nelson Goodman-style law-projectivists about counterfactuals,94 can make sense of possible worlds as merely possible (and not “as actual to their inhabitants as our world is to us,” a la David Lewis).

Another strand is the indeterminacy/vagueness of the similarity metric we are to use in assessing whether one possible world is closer to actuality than another. I have myself raised problems for Lewis’ four-fold similarity metric.95 Yet this very paper illustrates how in context similarity may not be so troublesome. For in the context of assessing abilities of persons to choose or do other than they did with an eye to assessing such persons’ responsibility, various features should plainly not vary, and only a limited set of features should be allowed to vary, as we judge similarity. Such contextualism reduces (even if it does eliminate) indeterminacy of the metric.

The third strand is a doubt that may linger even after the ontological and indeterminacy worries are allayed. As Tony Honore expressed this doubt (to Michael Smith’s possible world account of the counterfactuals unpacking ability to do otherwise), “what exactly does the

89 Model Penal Code ’4.02(1).
90 Model Penal Code ’2.09(1).
91 Model Penal Code ’210.3(1)(b).
94 Moore, Causation and Responsibility, pp. x, 390-392.
95 Id., pp. 385-390.
reference to possible worlds add?” In particular, once we eschew Lewis’ hoped-for possibilia as the truth-makers for counterfactuals, and use “possible worlds” talk as a way of talking about the states of affairs projected (by scientific law) to exist if certain other states of affairs exist, what is added by a mode of expression that we are not to take literally? The danger is that we substitute a fashionable patter for an unfashionable but more literal mode of expression.

A familiar feature of everyday life is the notion of a “near miss.” “A miss is as good as a mile” may be true in terms of an undesired outcome – for if all that matters is that it didn’t happen, well, in such cases it didn’t happen. But when our interests change, not all misses are created equal. Some of them – the near misses – give rise to expressions of relief, as in, “that was close!” We recognize that some small difference would have made a large difference in terms of producing an undesired outcome, and our relief lies in recognizing how easy it would have been – how little would have to have changed – for our lives to have gone a whole lot worse.

We thus intuitively appreciate the remoteness and closeness of what might have been. We don’t need to think there is an ontology of these states of affairs that might have been – “possibilia” – in order to use established scientific laws in our measurement of closeness. “Possible worlds” talk probably is dispensable; we probably could paraphrase to “law-projected states of affairs” talk, for example. But many find the unparaphrased possible worlds talk less cumbersome than the alternatives. And the phenomenon that talk of closeness of possible worlds is about – the near versus remote miss – is quite real.

IV. CONCLUSION

In closing, let me return to the basic building block of my analysis of volitional incapacity. This was the extensional equivalence of incapacity-on-an-occasion with certain counterfactuals about what would have happened in altered circumstances. Surely this move will cause some readers to protest:

“But no matter how close are the possible worlds in which actors like X would do what they should have done (not eat the cake), that was not the world in which X actually made his choice. In X’s actual world, the strength of his controlling desire and/or the non-stickiness of his executing intention, were such that that

---


desire was of lesser strength than his tempting desire, or that intention was not sticky enough to prevent vacillation. So in X’s actual world, he couldn’t have not eaten the cake – that choice and that act was fixed by the factors constituting the relative strength of his actual desires and by the factors constituting the stickiness of his actual intention.”

This is an expression of a kind of moral and metaphysical unease. When we move from “couldn’t have,” to “wouldn’t have,” we may seem to have eliminated both the excusing force of incapacity statements, and even incapacities themselves.

Those feeling the force of this worry should ask themselves what they mean by “could.” Seemingly in the foregoing they are giving the word its incompatibilist meaning: in the actual world, they think, X could not have refrained from eating the cake because X was sufficiently caused (by factors not under his control) to eat the cake. Yet this sense of “could,” leads either to the ultra-liberal absurdity that we are all excused, or to the desperate evasions of this conclusion I have examined in other papers.98

Suppose my imagined interlocutor eschews the incompatibilist reading of “could.” Is there still force to his worry, now expressed as a worry that in moving from what one can do in the actual world to what one would do in some possible world, the force of incapacity and thus of excuse is lost? We can alleviate these moral and metaphysical worries by not reducing what one could do to what one would do in changed circumstances; rather, we could say only that what one would do is merely a necessary concomitant to what one could do. Every ability, in other words, has its success conditions, even if the ability is not to be confused with the conditions of its verification. So the analysis needn’t eliminate “could” nor the sense of excuse built on it.

An alternative response would be to concede that the analysis leaves “could” behind, and that substituted for, “X couldn’t have done otherwise,” is, “It was too difficult for X to do otherwise.” Some criminal law theorists favor this move explicitly.99 Yet to my mind this concedes too much to the incompatibilist, namely, it allows them exclusive use of “could.” While Frankfurt-style compatibilists can afford to make such a concession (in light of their

---

99 E.g., Suzanne Uniacke, the commentator on this paper at its presentation in Scotland, favors this move.
rejection of the moral relevance of the principle of alternative possibilities entirely), my own kind of compatibilism cannot.\textsuperscript{100} I need to be analyzing “could,” not replacing it.

The intuitions behind the incompatibilist sense of “could” die hard. The other way in which they rear their head is in a kind of regress argument against the compatibilist (conditionalist) reading of “could.”\textsuperscript{101} The temptation is to regress “could” questions onto whatever condition is attached on the compatibilist analysis of “could.” Thus, to the conditional, “X would not have eaten the cake if his desire not to had been stronger,” one may be tempted to ask: “But could he have made his desire stronger?” Likewise, to the conditional, “X would not have eaten the cake if his intention not to do so had been stickier,” one may be tempted to ask: “But could he have made his intention stickier?” We are supposed to find intuitive a negative answer to these questions. And this would be right on the incompatibilist sense of “could.” The strength of our desires and the stickiness of our intentions are as caused as are any other macro-sized phenomena, and (in the sense of “uncaused”) we have no ability to strengthen desires or give backbone to intention.

This conclusion remains true if we use the compatibilist sense of “could” in this regress maneuver, viz, ask: if one tried to do so, would one affect the strength of desire or the stickiness of intention? Are these matters we can intentionally bring about, if we choose to do so? Some of my fellow compatibilists are tempted to waffle here: since James it is common to raise the possibility of strengthening desire or increasing stickiness of intention through indirect stratagems.\textsuperscript{102} Yet these in truth have limited efficacy. And in any case such response only delays the inevitable biting of the bullet that needs biting by the compatibilist. The truth is that strength of desire and stickiness of intention are by-and-large beyond our capacity to will directly or even to change by indirect stratagems; yet we are responsible just the same if that strength/stickiness happens to be such that with but little increase we would have done what we should, i.e., we could have done better. Compatibilism is at bottom the assertion that our responsibility is compatible with the existence of moral luck in who we are, even though who we are is often not within our control.

\textsuperscript{100} Moore, “Compatibilism(s) for Neuroscientists.”
\textsuperscript{101} See \textit{id}.
\textsuperscript{102} E.g., Jay Wallace, “Addiction as Defect of the Will,” pp. 635-636.