Pace Law Review

Volume 36
Issue 3 *Spring 2016*

Article 1

August 2016

Retributivism, Agency, and the Voluntary Act Requirement

Christopher P. Taggart
Harvard Law School

Recommended Citation
Christopher P. Taggart, *Retributivism, Agency, and the Voluntary Act Requirement*, 36 Pace L. Rev. 645 (2016)
DOI: https://doi.org/10.58948/2331-3528.1926
Available at: https://digitalcommons.pace.edu/plr/vol36/iss3/1

Retributivism, Agency, and the Voluntary Act Requirement

Christopher P. Taggart*

Abstract

The Voluntary Act Requirement ("VAR") is the fundamental predicate for imposing legal punishment. Punishing solely on the basis of evil thoughts or a villainous character is impermissible. The VAR also embodies the notion that we must not punish someone for conduct over which she lacked sufficient control. But why not punish someone for conduct that was not within her control? One answer is retributivist—it would be unjust to do so because that defendant could not have been morally responsible for, and therefore could not deserve punishment for, what she did. Agent causalism is a contentious view about how criminal defendants voluntarily act according to which the defendants themselves cause their free, morally responsible actions, as opposed to events or states of affairs involving them, their brains, their circumstances, and so forth. This article argues that for retributivist justifications of the VAR to be plausible, agent causalism must be true. Agent causalism might be false, and if it is, then retributivism could not play any role in justifying our fundamental legal precondition for ever imposing any criminal liability upon anyone. This article does not argue that agent causalism is false, however. It elaborates and renders plausible an agent causalist position, and it shows how that position could handle types of cases that notoriously pose challenges to the VAR—cases involving complex unconscious conduct, cases involving crimes of omission, and cases involving habitual conduct.

I. Introduction

* Lecturer on Law, Harvard Law School. I wish to thank Palma Paciocco and Mike Materni for very helpful comments on earlier drafts.

Our criminal legal system requires that someone act voluntarily before legal punishment is imposed upon her. This idea is fundamental to the state ever being justified in punishing anyone, and it is enshrined in our law in the form of the Voluntary Act Requirement ("VAR"). Indeed, the American Law Institute ("ALI") deems the VAR "the fundamental predicate for all criminal liability . . . ."[1] According to the VAR, "the guilt of the defendant [must] be based upon conduct, and that conduct must include a voluntary act or an omission to perform an act of which the defendant [is] physically capable."[2]

Because the VAR is the fundamental precondition of any criminal liability, justifying the VAR is an important element of justifying our system of criminal law and our institution of criminal punishment. At stake is justifying not just coercive, painful government intrusion into the lives of criminals. Punishment hurts criminals *on purpose*. Ideally, we should have a good reason justifying every aspect of doing *that*. If criminal law's inclusion of the VAR is not defensible, then any attempt systematically to justify imposing punishment on anyone under our system would be woefully incomplete.

Utilitarian[3] and retributive theories of punishment are the two main avenues by which such explanations and justifications proceed.[4] This article elaborates a relationship between

1. MODEL PENAL CODE § 2.01 explanatory note on subsection (1) (AM. LAW INST., Official Draft and Revised Comments 1985).

2. *Id.* Put differently:

> A conviction of a defendant for crime C is justified only if (1) There is a voluntary act, the performance of which is necessary for C's occurrence (given the statutory definition of C) and (2) the defendant has been shown (typically, beyond a reasonable doubt) to have performed such a voluntary act.

Gideon Yaffe, *The Voluntary Act Requirement*, *in* THE ROUTLEDGE COMPANION TO THE PHILOSOPHY OF LAW 174, 174 (Andrei Marmor ed., 2012).

3. I use *utilitarian* broadly to include any consequentialist, welfarist approach to normative analysis. Included are welfarist views according to which the analyst does not simply sum up the values of individual utilities to determine the value of a situation featuring those individuals.

4. *See* C. L. TEN, CRIME, GUILT, AND PUNISHMENT 3 (Oxford Univ. Press 1987) ("The philosophical debate on punishment has been dominated by two main types of theories of punishment, the utilitarian theory and the retributive theory.").

retributivism and the VAR. There is a contentious[5] view about how criminal defendants, and human beings generally, voluntarily act, called agent causalism. Roughly, agent causalism is the view that when we voluntarily act and are morally responsible for what we do, we are the causes of our conduct, as opposed to events or states of affairs, internal or external to us, that deterministically cause our conduct. For example, if Fred assaults someone voluntarily and is morally responsible for doing so, then Fred causes his conduct—not states of or events in Fred's brain, but Fred himself. As will be elaborated, agent causalism has emerged from attempts to solve what has been called the problem of free will.

This article's main thesis is that for retributivist justifications of the VAR to be plausible, agent causalism must be true. This thesis has a significant implication. Because agent causalism is contentious, it might be false.[6] And if agent causalism is false, then, according to this article's thesis, it would be implausible that retributivism could play any role in substantiating the fundamental legal precondition of ever imposing any criminal liability upon anyone.

This article's point of departure is the VAR as it is codified at § 2.01 of the Model Penal Code ("MPC"). Therefore, it is important to address one of the ALI's official comments on MPC § 2.01 up front:

> The term "voluntary" as used in [§ 2.01] does not inject into the criminal law questions about determinism and free will. Rather, it focuses upon conduct that is within the control of the actor. There is sufficient difference between ordinary human activity and a reflex or a convulsion to make it desirable that they be distinguished for purposes of criminal responsibility by a term like

5. By *contentious*, I do not mean merely *likely to cause controversy*. By a *contentious* view, I also mean that reasonable, informed people disagree about the view. For a view to be *contentious*, it must be at least coherent and somewhat plausible.

6. Although I am inclined to think that agent causalism is true, this paper will try to establish only that it is coherent and plausible.

"voluntary."[7]

This comment seems to rule out the idea that a normative theory's ability to justify the VAR could turn on the relationship between that theory, retributivism, and a view that is motivated by responding to questions about determinism and free will, agent causalism. The comment also clarifies that the rationale behind the VAR focuses on whether an actor has control over her conduct. Why does it matter whether the defendant was in control of what she did when we decide whether she is criminally responsible? Why not punish someone for conduct that was not within her control? Perhaps the most natural, but by no means the only, answer to this last question is that it would be unjust to do so because that defendant could not have been morally responsible for what she did.[8]

Some of the questions that arise in debates about how to solve the problem of free will concern whether an actor can ever be morally responsible for her conduct given the degree or type of control over her conduct that she has or lacks. Many care about the problem of free will because: (a) they want to have enough of the right kind of control over what they do to be morally responsible for what they do, at least sometimes, but (b)

7. MODEL PENAL CODE § 2.01 cmt. 1 (AM. LAW INST., Official Draft and Revised Comments 1985).

8. This answer is natural in part because a paradigmatic instance of criminal punishment "must be of an actual or supposed offender for his offense." H. L. A. Hart, *Prolegomenon to the Principles of Punishment*, *in* PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 5 (Oxford Univ. Press 1967). Along similar lines:

> I bring myself within the reach of the criminal law only when I act: only when thought and intention are given active embodiment in conduct which engages with the world, and which may thus impinge on the rights and interests that the criminal law aims to protect. It also . . . seems appropriate because we can surely be held culpably responsible only for what is within our control; and, once we move beyond the realm of (mere) thought, it is our actions that we paradigmatically control.

R. A. Duff, *Acting, Trying, and Criminal Liability*, *in* ACTION AND VALUE IN CRIMINAL LAW 75, 78 (Stephen Shute et al. eds., 1993) [hereinafter Duff, *Acting, Trying and Criminal Liability*].

causal determinism[9] seems to threaten that sort of control. If they lack the requisite control, then they lack moral responsibility.

The ALI's comment to MPC § 2.01 suggests that: (i) issues surrounding the problem of free will have no bearing on the VAR, but (ii) issues surrounding an actor's control are of central importance to the VAR. Therefore, it would seem that whatever sense of a defendant's *control* over her conduct the ALI has in mind must not be connected with whether that defendant is ever morally responsible for what she does.[10] If what the ALI had in mind by an actor's *control* had anything to do with her moral responsibility, then, the ALI's protestations notwithstanding, the ALI would be *injecting into criminal law* questions about determinism and free will.

Some questions about determinism and free will, the ones that animate the problem of free will, automatically get *injected* into any discussion that turns on whether an agent has the control requisite for moral responsibility. This *injection* follows from: (1) what the problem of free will is about and one central reason why it is significant, that is, the connection between moral responsibility and control, and (2) the fact that some questions about determinism and free will arise in connection with that problem. If control and moral responsibility are connected as they seem to be in many discussions of determinism and free will, then the ALI is injecting questions about determinism and free will into criminal law even though it claims not to be.

Assume, however, that the ALI is not injecting questions about determinism and free will into criminal law. More specifically, assume that: (a) the ALI's rationale behind the VAR

9. Causal determinism is the view that "the past and the laws of nature together determine, at every moment, a unique future. . . ." Peter van Inwagen, *How to Think about the Problem of Free Will*, 12 J. ETHICS 327, 330 (2008) [hereinafter van Inwagen, *How to Think*].

10. I wish to thank Palma Paciocco and Mike Materni for pointing out that the ALI might be assuming that questions about determinism and free will are not injected into the criminal law by the VAR because the problem of free will has a solution, according to which we are often free and morally responsible for what we do. In other words, the ALI might be taking it for granted that a defendant must be morally responsible for what she does if she is to be punished for her conduct and that in most cases in which she has control over her conduct she is morally responsible for what she does.

turns on an agent's control over her conduct but (b) an agent's control over her conduct has nothing to do with her moral responsibility for her conduct. That is, assume that an agent's moral responsibility for her conduct has nothing to do with the ALI's rationale behind the fundamental predicate for criminal liability.[11] Perhaps the important thing for the ALI is that sufficiently-controlled conduct is required to justify imposing legal punishment, as opposed to: (i) some kind of non-conduct, such as a *status*[12] or *mere thought*,[13] over which the defendant has less control or (ii) behavior such as a tic or seemingly goal-oriented yet unconscious behavior,[14] over which the defendant

11. To emphasize, this is just a working assumption. It is not my goal to offer the best interpretation of what the ALI's view is.

12. *See, e.g.*, Robinson v. California, 370 U.S. 660 (1962). In *Robinson*, the Court held that "a state law which imprisons a person [addicted to narcotics] as a criminal, even though he has never touched any narcotic drug within the State or been guilty of any irregular behavior there, inflicts a cruel and unusual punishment in violation of the Fourteenth Amendment." *Id.* at 667. In reaching this holding, the Court stressed that the statute at issue made "the 'status' of narcotic addiction a criminal offense, for which the offender may be prosecuted 'at any time before he reforms.'" *Id.* at 666.

13. There are rationales for not criminalizing thoughts that do not emphasize the agent's degree of control over her own thoughts. For example, even if the agent had the same type and degree of control over her thoughts that she had over her voluntary actions, it would be impermissible to criminalize thoughts because of her right to freedom of thought. "There is something objectionable about criminalizing thoughts alone. Prohibitions on thoughts are intrusive violations of privacy, efforts at mind control, and inconsistent with the goals and role of a liberal state." Yaffe, *supra* note 2, at 175.

14. In People v. Newton, 87 Cal. Rptr. 394 (Ct. App. 1970), the defendant appealed his conviction for voluntary manslaughter, arguing that during the time that he shot his victim, he was not conscious because he himself was in an altered state caused by having been shot in the abdomen. The conviction was reversed. *Id.* at 415. An expert witness testified that the defendant could have been in a "reflex shock condition" in which the defendant unconsciously engaged in complex goal-oriented behavior usually indicative of conscious action. *Id.* at 403. Central to the reversal was that:

> The difference between . . . diminished capacity and unconsciousness . . . is one of degree only: where the former provides a "partial defense" by negating a specific mental state essential to a particular crime, the latter is a "complete defense" because it negates capacity to commit any crime at all.

Id. at 405-06.

has less control. The defendant's moral responsibility for her status, thoughts, behavior, or controlled conduct is entirely beside the point. Even if the ALI thought that moral responsibility was irrelevant to the fundamental predicate of criminal liability, we could, and should, ask whether such a position is defensible. Even though its views are worthy of serious consideration, the ALI, of course, is not a *primary legal* or *moral* authority.

Of the two main competing theories—utilitarianism and retributivism—the idea that moral responsibility is irrelevant to the VAR is more at home with utilitarianism. Utilitarian theories do not center on the evaluation of actors as retributive theories do.[15] Whether an actor is morally responsible for her conduct is central to how retributive justificatory reasons work, if they work at all. But for the utilitarian, if: (i) an actor is not morally responsible for her conduct but (ii) nonetheless her conduct can be influenced by influencing her, then the VAR might be substantiated by appealing to, say, punishment's deterrent effects. The basic utilitarian idea would be that for a criminal defendant, or anyone, to be deterred from acting a certain way, she must be in sufficient control of her conduct, even if she bears no moral responsibility for her conduct.[16]

15. *See infra* Part II for an elaboration on the focus of utilitarian theories on evaluating actions on the basis of their consequences instead of focusing on evaluating actors.

16. Of course, this sort of utilitarian approach has its own challenges to overcome. For example, a utilitarian might try to justify the VAR by arguing that:

> Those who only wish and fantasize criminal acts, but don't actually do them, aren't dangerous; those whose (involuntary) clumsiness cause[s] others harm aren't deterrable; etc. Yet it is not obvious that these generalizations hold. Mightn't 'accident-prone' individuals be dangerous, and thus subject to preventative detention on utilitarian grounds? Mightn't such classes of individuals be somewhat deterrable, a least to the extent that they could take some precautions against their dangerous tendencies? And even if they themselves are not deterrable, mightn't the criminal law gain an increment of general deterrence by making such persons liable anyway, because then those voluntarily causing harm will know that there is no possibility of *pretending* to have involuntarily caused it?

The retributivist does not rely on this idea. Instead, the basic retributivist idea is that someone should be punished because she deserves it. Does the criminal defendant's desert warrant punishment, and if so, how much? Any retributivist attempt to justify the VAR must address how the VAR helps insulate those who are not morally responsible for their conduct, and therefore lack desert, from criminal liability.[17]

Consider the following: (i) a necessary condition for an actor's desert is that she conduct herself in some way and be morally responsible for that conduct; (ii) a necessary condition for an actor to be morally responsible for her conduct is that she have the right sort of control over her conduct; (iii) a necessary condition for an actor to have the right sort of control over her conduct is that agent causalism be true; therefore, (iv) a necessary condition for an actor's desert is that agent causalism be true. Since a criminal defendant's desert is the central idea of retributivist justifications, including any such justification of the VAR, for retributivist justifications of the VAR to be plausible, agent causalism must be true. That is the kernel of this article's argument for its main thesis.

After presenting the argument for the main thesis, I shall spend considerable space examining whether agent causalism is coherent and plausible. As previously explained, the significance of the main thesis turns in part on whether agent causalism is contentious, and to be contentious agent causalism must be coherent and at least somewhat plausible. I shall also apply an agent-causal retributivist approach to justifying the VAR to three categories of non-paradigmatic cases. The goal will

MICHAEL S. MOORE, ACT AND CRIME: THE PHILOSOPHY OF ACTION AND ITS IMPLICATION FOR CRIMINAL LAW 47 1993) [hereinafter MOORE, ACT AND CRIME].

17. I am not suggesting that according to retributivist theories the VAR must shoulder the entire burden of shielding those not morally responsible from criminal liability. For example, a retributivist might think that much of that burden is borne by the MPC's culpability requirements or by Article 4. For example, Model Penal Code § 4.01(1) absolves an actor of criminal liability for his conduct if "at the time of such conduct as a result of mental disease or defect he lacks substantial capacity either to appreciate the criminality . . . of his conduct or to conform his conduct to the requirements of law." Also, note that according to standard retributivist theories, desert is not merely necessary for punishment. It is also sufficient. In connection with justifying the VAR, however, the part of retributivism that is particularly relevant is the view that desert is necessary for punishment.

be to bolster the credibility of agent-causal retributivism by showing how it yields defensible results even in hard cases. More specifically, the remainder of this paper proceeds as follows:

In Part II, I explain that a defensible retributivist theory requires that for someone to deserve legal punishment, she must conduct herself in some way and be morally responsible for that conduct. I do this by distinguishing utilitarian theories from retributivist theories and then elaborating the key retributivist notion of desert in light of those distinctions. In Part III, I examine what sort of control over her own conduct would be necessary for someone to be morally responsible for her conduct, and therefore to deserve legal punishment for it. I do this by discussing two related problems concerning moral responsibility and control over what one does—the problem of moral luck and the problem of free will. In Part IV, I present agent causalism as a view that affords actors the sort of control necessary for moral responsibility. To substantiate my claim that agent causalism is contentious, I argue that agent causalism is coherent and at least somewhat plausible, even though I do not attempt a thorough defense of agent causalism. In Part V, I continue to argue that agent causalism is plausible. Drawing heavily on the work of others, I sketch a picture of how agents fit into voluntary actions resulting from practical deliberation to explain how, according to agent causalism, the way that agents voluntarily act might be responsive to practical reasons. In Part VI, I bolster the contention that agent causalism is plausible by showing how it might aid the retributivist in regard to three sorts of cases in which the VAR is implicated—cases involving complex unconscious conduct, cases involving crimes of omission, and cases involving habitual conduct. I conclude with some brief summarizing remarks.

II. Desert Requires Moral Responsibility for Voluntary Action

As mentioned above, there are two dominant types of justifications of legal punishment—utilitarian and retributivist. A normatively important question regarding punishment is: "what justifies the state in inflicting hard treatment on people for their supposed or claimed wrongdoing with the intention that

that treatment cause the supposed or claimed wrongdoer to suffer?"[18] When the state punishes, it purposely inflicts suffering upon the defendant. In a case where a defendant is found not guilty of murder by reason of insanity, he may be confined to an institution to protect the public. And his confinement might cause him to suffer. But the point of confining him is not to inflict suffering upon him.[19] Because legal punishment, in contrast, purposely, not just knowingly, inflicts suffering, the call for its justification is especially exigent.

A. *Utilitarian Theories of Punishment*

Utilitarian[20] theories of punishment center on the effects of punitive practices and decisions on the well-being of individuals in society—criminal defendants included.[21] They are

18. Mitchell N. Berman, *The Justification of Punishment*, *in* THE ROUTLEDGE COMPANION TO THE PHILOSOPHY OF LAW 141, 143 (Andrei Marmor ed., 2012).

19. *See* DAVID BOONIN, THE PROBLEM OF PUNISHMENT 13 (Cambridge Univ. Press 2008) ("In [confining him], the state recognizes that its action will seriously harm the [defendant], but harming him is not its intention. Its intention is merely to protect the public, and it would lock him up even if this did not harm him.").

20. Technically, utilitarianism is committed to specific ways of amalgamating utilities when assessing the value of a state of affairs. The utilitarian either takes the sum of individual utilities (classical utilitarianism) or the average of individual utilities (average utilitarianism) in computing a numerical representation of the value of a situation—that situation's *amount of social welfare*. I intend my claims about a utilitarian approach to justifying punishment to carry over to any *welfarist* approach that is committed to consequentialism. Welfarism is the view that the only features of a state of affairs that determine the state's intrinsic value are, collectively, the state's utility information (i.e., a pairing of each individual in a situation with her utility in that situation).

21. The central utilitarian idea has been expressed in a number of ways. *See* JOSHUA DRESSLER, CRIMINAL LAW 14 (5th ed. 2009) ("according to classical utilitarianism . . . the purpose of all laws is to maximize the net happiness of society. Laws should be used to exclude, as far as possible, all painful and unpleasant events . . . . [B]oth crime and punishment are unpleasant . . . . [T]he pain inflicted by punishment is justifiable if, but only if, it is expected to result in a reduction in the pain of crime that would otherwise occur."); TEN, *supra* note 4, at 3 ("The utilitarian theory justifies punishment solely in terms of its beneficial effects or consequences . . . . [U]ltimately the only morally significant features of an act are the good and bad consequences produced by it. A right act is that which, among the available alternatives, produces the best

consequentialist theories. According to consequentialism, only the consequences of implementing feasible options are relevant to what choices morally ought to be made.[22] "[C]onsequentialism is the doctrine that the moral value of any action always lies in its consequences, and that it is by reference to their consequences that actions . . . are to be justified if they are to be justified at all."[23] A consequentialist does not merely aver that consequences are of primary moral importance. She claims that only consequences are morally relevant to choice. Thus, a non-consequentialist might consistently think that the consequences of feasible alternatives are always very important moral considerations. As John Rawls emphasizes, it is a mistake to think that non-consequentialist theories "characterize the rightness of institutions and acts independently from their consequences. All ethical doctrines worth our attention take

consequences."); LLOYD L.WEINREB, CRIMINAL LAW: CASES, COMMENT, QUESTIONS 327 (7th ed. 2003) ("punishment is justified by its utility, the good that it does, not necessarily for the criminal himself but for the community.").

22. It is possible for a theorist to be a consequentialist when it comes to justifying punishment without being a consequentialist tout court:

> Because it is customary to classify moral theories . . . as either consequentialist or deontological, it is tempting to suppose that consequentialist theories of punishment must be committed to a consequentialist ethic . . . . However, the mapping of consequentialist theories of punishment onto consequentialist moral theories is too facile . . . . Consequentialism in punishment theory is a view regarding how the intentional infliction of suffering for wrongdoing can be morally justified; it is not a view about value or right action more generally.

Berman, *supra* note 18, at 144.

Of course, anyone who is a consequentialist vis-à-vis punishment but who denies consequentialism with respect to the evaluation of other important social choices presumably has reasons for the discontinuous nature of her approach. And it would be fair to ask such a theorist what those reasons are. For example, why be a consequentialist when it comes to justifying legal punishment but not be a consequentialist when it comes to, say, justifying one particular redistributive tax-and-transfer regime over others? Putting this aside, since I am discussing only theories of legal punishment, I shall assume that the possibility of a *fair weather* consequentialist does not impugn the details of my characterization of a utilitarian theory of punishment as a type of consequentialist theory of punishment.

23. Bernard Williams, *A Critique of Utilitarianism*, *in* UTILITARIANISM: FOR AND AGAINST 75, 79 (J.J.C. Smart & Bernard Williams ed., 1973).

consequences into account in judging rightness. One which did not would simply be irrational, crazy."[24]

Of particular significance for this paper, "a central idea of consequentialism [and therefore of utilitarianism] is that the only kind of thing that has intrinsic value is a state of affairs, and that anything else that has value has it because it conduces to some intrinsically valuable state of affairs."[25] For a consequentialist, consequences are all that ever ultimately matter morally, and consequences are states of affairs.[26] The deontic status[27] of a social, or individual, choice depends on the comparative intrinsic values of the states of affairs that would be brought about by the options that are feasible for society, or the individual, at the time of choice. For this reason, the criminal defendant is not the central object of normative assessment for utilitarian justifications of punishment. The utilitarian takes an *ex ante* point of view—the social choices (a) to adopt a particular system of criminal punishment and (b) to impose, under that system, a certain amount of legal punishment upon a particular defendant are to be justified by the consequences of doing so. If the consequences of such choices are better than those of any feasible alternatives, then we should make those particular choices—our legally punishing in that way is justified.

An obvious potentially beneficial effect of legal punishment is crime reduction. Accordingly, it is common for utilitarian analysts to focus on various ways that punishment reduces crime when they offer justifications for legal punishment.

24. JOHN RAWLS, A THEORY OF JUSTICE 30 (Harvard Univ. Press 1971). This is not to deny that a non-consequentialist of an extreme sort might think that the consequences of feasible options are always irrelevant to moral choice. Such a view is one possible type of non-consequentialist view. (In my opinion, such an extreme form of non-consequentialism is very implausible.) But to reiterate the main point, it is a misunderstanding to think that non-consequentialists characteristically do not consider consequences to be important, morally relevant factors when making social choices.

25. Williams, *supra* note 23, at 83.

26. Informally, one might say that a consequence is a type of situation—a situation that results from or is the outcome of an action or choice—such as the choice to imprison Fred for five years or the choice not to impose criminal liability upon Linda or the choice to abolish all *strict liability* crimes.

27. For example, *morally permissible*, *morally forbidden*, and *morally required*.

Publically punishing Fred for committing a crime is justified because it reduces crime by scaring people into not committing it (general deterrence); punishing Fred for committing a crime is justified because the horrible experience of being punished will scare Fred into not committing future crimes (specific deterrence); punishing Fred for committing a crime is justified because punishing him improves his character so that he will not commit future crimes (rehabilitation); punishing Fred for committing a crime by incarcerating him is justified because incarcerating him prevents him from committing future crimes, at least for as long as he remains incarcerated.[28]

Notice that the thing that the utilitarian promotes in the previous paragraph is always a situation with less crime going on in it. Of course, utilitarians care, often a lot, about certain features of agents—for example, whether Fred is dangerous, how well-off Fred or anyone else would be under various circumstances, and so forth. And a utilitarian might even think, in some derivative sense, that such features have moral or ethical[29] significance. But the only type of thing that has any intrinsic moral/ethical value for a utilitarian is a situation. For the utilitarian, there are primarily two types of things that get morally/ethically assessed—choices, actions, and outcomes. The assessment of a choice, as permissible, impermissible, etc., depends on a prior assessment of the intrinsic moral/ethical value of its outcome. And for a utilitarian, the moral/ethical value of an outcome is a function solely of the utilities of the individuals, including the criminal offenders, *in*[30] that outcome.

28. Of course, crime reduction need not be the only potentially welfare-enhancing consequence that a utilitarian theorist of legal punishment focuses on. For example, if we assume a *preference-satisfaction* interpretation of individual utility, then, if enough persons have a stable preference that offenders receive what might be considered their *just deserts*, then legally punishing in a certain way might significantly increase individual utilities, and therefore increase social welfare. *See, e.g.*, A. Mitchell Polinsky & Steven Shavell, *The Fairness of Sanctions: Some Implications for Optimal Enforcement Policy*, 2 AM. L. & ECON. REV. 223 (2000).

29. Some may try to distinguish the concepts expressed by terms such as *moral*, *morality*, and *morally* from the concepts expressed by terms such as *ethical*, *ethics*, and *ethically*. This article does not draw such distinctions.

30. To say that Fred is *in* a state of affairs (or situation, or outcome) is to say that if that state of affairs were actual, then Fred would exist. And Fred's utility *in* an outcome refers to how well-off Fred would be if that outcome were to come to pass. This article will not consider how defensible a standard

B. *Retributivist Theories of Punishment*

Retributivists, in contrast, do not primarily focus on good or bad consequences. For a standard sort of retributivist, the moral assessment of agents—the criminal defendants themselves—plays a primary role. As explained, even if a utilitarian morally evaluates agents, that evaluation is secondary to the intrinsic moral/ethical value of realized states of affairs and the derived evaluation of choices that lead to them. The moral feature of an agent that retributivist theories focus on is her desert.[31] And an agent's desert is conceptually connected to voluntary wrongful conduct, since "retributive theories of punishment . . . maintain that punishment is justified because the offender has voluntarily committed a morally wrong act."[32]

utilitarian or welfare-economic view of the nature of individual well-being is.

31. TEN, *supra* note 4, at 46 ("Contemporary retributivists treat the notion of desert as central to the retributive theory, punishment being justified in terms of the desert of the offender.").

32. *Id.* The central retributivist idea has been expressed in a number of ways. *See id.* at 5 ("Retributivists regard the offender's wrongdoing as deserving of punishment, and the amount of punishment should be proportionate to the extent of wrongdoing. The offender's desert, and not the beneficial consequences of punishment, is what justifies punishment"); Anthony Duff, *Legal Punishment, in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY (2001) [hereinafter Duff, *Legal Punishment*], http://plato.stanford.edu/entries/legal-punishment/ ("The guilty, those who commit criminal offences, deserve to be punished: which is to say . . . not merely that we must not punish the innocent, or punish the guilty more than they deserve, but that we should punish the guilty, to the extent that they deserve: penal desert constitutes not just a necessary, but an in principle sufficient reason for punishment."); MICHAEL S. MOORE, PLACING BLAME: A THEORY OF THE CRIMINAL LAW 153 (Oxford Univ. Press 1997) [hereinafter MOORE, PLACING BLAME] ("[R]etributivism is the view that we ought to punish offenders because and only because they deserve to be punished. Punishment is justified, for a retributivist, solely by the fact that those receiving it deserve it."); WEINREB, *supra* note 21, at 327 ("[P]unishment is retribution for the wrong done by the criminal; it is retrospective, a requirement of justice justified directly and completely by the past conduct of the person punished . . . . Not only does retribution justify punishment; it prohibits a relaxation of punishment in order to accomplish some social good.").

One of the better known illustrations of the "non-relaxation" aspect of retributivism comes from Kant:

> [W]hoever has committed murder, must die . . . . Even if a civil society resolved to dissolve itself with the consent of all

Retributivism is not just an academic theory. It is operative in legal opinions as well. To offer a couple of high-profile examples: In *Enmund v. Florida*,[33] Earl Enmund was sentenced to death as an accomplice to felony murder.[34] Enmund had not killed anyone.[35] Nor did the original plan include killing anyone.[36] Enmund appealed his sentence to the Court, which reversed, barring Florida from executing him.[37] In reversing Enmund's death sentence, the Court reasoned:

> Here the robbers did commit murder; but they were subjected to the death penalty only because they killed as well as robbed. The question before us is not the disproportionality of death as a penalty for murder, but rather the validity of capital punishment for Enmund's own conduct. The focus must be on *his* culpability, not on that of those who committed the robbery and shot the victims, for we insist on "individualized consideration as a constitutional requirement in imposing the death sentence" . . . . Enmund did not kill or intend to kill and thus his culpability is plainly different from that of the robbers who killed; yet the State treated them alike and attributed to Enmund the culpability of those who killed the [victims]. This was impermissible under the Eighth Amendment.[38]

> its members . . . the last murderer lying in prison ought to be executed before the resolution was carried out. This ought to be done in order that every one may realize the desert of his deeds, and that blood-guiltiness may not remain upon the people; for otherwise they might all be regarded as participators in the murder as a public violation of justice.

IMMANUEL KANT, THE SCIENCE OF RIGHT, The Right of Punishing and of Pardoning § E(I) (W. Hastie trans., 2003) (1790), http://xet.es/books/Kant/The%20Science%20of%20Right%20Kant.pdf.

33. Enmund v. Florida, 458 U.S. 782 (1982).
34. The underlying felony was an armed robbery. *Id.* at 784-85.
35. *Id.* at 784.
36. *See generally id.*
37. *Id.* at 801.
38. *Id.* at 798 (citations omitted).

Although Enmund was guilty of felony murder, it was impermissible to execute him for that crime because, unlike the others who shot and killed the crime victims, Enmund's desert did not warrant that severe a punishment. His culpability was insufficient.

In *Atkins v. Virginia*,[39] Daryl Atkins was sentenced to death for capital murder and appealed his sentence to the Court, which reversed.[40] The Court agreed with Atkins's argument; because he was mentally retarded, he could not be lawfully sentenced to death.[41] The Court took note of how a consensus among states to disallow the execution of mentally retarded defendants "unquestionably reflect[ed] widespread judgment about the relative culpability of mentally retarded offenders, and the relationship between mental retardation and the penological purposes served by the death penalty."[42] And the Court agreed that, while mentally retarded defendants' "deficiencies do not warrant an exemption from criminal sanctions . . . they do diminish their personal culpability."[43] Further:

> With respect to retribution—the interest in seeing that the offender gets his "just deserts"—the severity of the appropriate punishment necessarily depends on the culpability of the offender. . . . [O]ur jurisprudence has consistently confined the imposition of the death penalty to a narrow category of the most serious crimes. . . . If the culpability of the average murderer is insufficient to justify the most extreme sanction available to the State, the lesser culpability of the mentally retarded offender surely does not merit that form of retribution.[44]

Thus, one significant reason why the Court forbade the

39. Atkins v. Virginia, 536 U.S. 304 (2002).
40. *Id.* at 321.
41. *See generally id.*
42. *Id.* at 317.
43. *Id.* at 318.
44. *Id.* at 319.

execution of Atkins was that Atkins could not have had the culpability necessary to deserve death for his crime.

As previously mentioned, the retributivist is interested primarily in a defendant's desert. Before elaborating the notion of desert more fully, I should point out that what I take to be a *standard* retributivist theory is not a form of consequentialism that focuses on minimizing the number of persons who fail to be punished as they deserve to be.[45] Unlike utilitarianism's moral rationality, the *standard* moral rationality of retributivism is neither minimizing nor maximizing. To illustrate one type of consequentialist theory that I wish to distinguish from what I take to be a more standard retributivist view, I would like to consider a particular *welfare-economic* critique of setting the level of punishment for a certain crime on the basis of the *retributively fair* level.[46] The gist of the critique is that the retributivist adopts a way of assessing the outcomes of competing legal and policy choices that tends to recommend choices leading to inferior outcomes, even as those outcomes are evaluated by retributivist lights. The consequentialist retributivist favorably ranks outcomes in which punishments are properly proportioned to the retributively fair level to *fit* the crimes committed. Such outcomes are rated higher than other outcomes in which punishments are more severe and widely publicized and thereby manage to scare potential criminals enough that no crimes are committed. That is, when the retributively fair punishment is selected, some undeterred people will commit crimes, and many will get away with them. Such individuals go unpunished and are therefore treated *unfairly*—they do not get what they deserve:

45. According to this type of consequentialist theory, if someone deserves no punishment and is not punished at all, then she is punished as she deserves to be (viz., not at all). So this form of retributive consequentialism would consider a society in which there was no crime and no punishment to be ideally minimizing (though there would be other ideally minimizing possibilities). Also, to fail to be punished as one deserves to be, one is either punished more severely or less severely than one deserves to be. For example, someone who deserves a little punishment but is not punished at all is not punished as she deserves to be.

46. *See* LOUIS KAPLOW & STEVEN SHAVELL, FAIRNESS VERSUS WELFARE 320–29 (Harv Univ. Press 2002).

> It is peculiar . . . for retributivists to insist that the sanction should not exceed the fair ideal . . . regardless of how much unfairness results with regard to those who go scot-free . . . . [U]nder the unfair sanction [that deters], no one . . . receives unfair treatment. Therefore, when one considers the unfairness surrounding the punishment of all the criminals who commit the wrongful act when the sanction is [fair], one should be troubled. The [retributive] fairness view, on its own terms, seems erroneously constrained as it only considers the [] individuals who are caught and ignores . . . [those] who are not.[47]

This critique may pose a problem for a type of retributivist who emphasizes the comparative evaluation of outcomes on the basis, a least in part, of how much *unjust-because-undeserved* punishment is realized in the outcomes being compared.[48] But such a retributivist does not hold what I take a more standard sort of retributivist theory to be, especially in regard to how to justify the imposition of a certain amount of punishment upon a particular individual on the basis of what he or she has done. What I understand to be a more standard retributivist view does not focus on maximizing the value of outcomes in the way the consequentialist retributivist view targeted by the critique does.

A more standard retributivist view[49] also justifies

47. *Id.* at 325.

48. As explained, to be saddled with the problem that the critique poses, the consequentialist retributivist would also need to rank situations higher when the general levels of punishment are set to the retributively fair level. This sort of retributivist might be able to escape the charge that her position is erroneous on its own terms if she is prepared to explain how two different properties of outcomes—(1) the fairness of general levels of punishment and (2) the amount of *unjust because undeserved punishment*—are to be traded off against one another under the chosen retributive social welfare function. If she can do this, then it might turn out that her theory consistently provides a high ranking to situations in which there is a lot of *unjust because undeserved punishment*, as long as such situations feature general criminal penalties that are *extremely fair*. Of course, the price that the retributivist might need to pay to take this tack is that her weighting of the two different (retributive) fairness-based properties under the proposed social welfare function would be extremely implausible.

49. From here on, only a more standard non-consequentialist view will be

punishment on the basis of desert, but in a different way. When assessing whether legally to punish a defendant and how much to punish her, the retributivist focuses on whether she, the agent, deserves punishment and if so, then how much. Although judging what an agent deserves is a way of judging the agent herself, what the agent does also plays an indispensable role:

> If a person is deserving of some sort of treatment, he must, necessarily, be so *in virtue of* some possessed characteristic or prior activity. It is because no one can deserve anything unless there is some basis or ostensible occasion for the desert that judgments of desert carry with them a commitment to the giving of reasons. One cannot say, for example, that Jones deserves gratitude although he has done "nothing in particular." If a person says that Jones deserves gratitude, then he must be prepared to answer the question "For what?" Of course, he may not know the basis of Jones's desert, but if he denies that there is any basis, then he has forfeited his right to use the terminology of desert. He can still say that we *ought* to treat Jones well for "no reason in particular" of simply "for the sake of being nice," but it is absurd to say that Jones *deserves* good treatment for no reason in particular. Desert without a basis is simply not desert.[50]

Assertions of desert have an implicit structure: "'*S* deserves *X* in virtue of *F*,' where *S* is a person, *X* is a mode of treatment, and *F* [is] some fact about *S*. . . ."[51] Further, if *X* is legal punishment, then *F*, the fact about *S*, must be a fact about something *S* did.[52] *F* cannot be a fact about *S*'s status or about *S*'s mere thoughts or feelings. The government should not

considered.

50. Joel Feinberg, *Justice and Personal Desert*, *in* DOING AND DESERVING: ESSAYS IN THE THEORY OF RESPONSIBILITY 55, 58 (Princeton Univ. Press 1970) [hereinafter Feinberg, *Justice and Personal Desert*].

51. *Id.* at 61.

52. Here I am glossing over the act/omission distinction.

punish people merely because they have villainous characters or evil thoughts.[53] To say that A deserves a certain amount of legal punishment is to assess A herself, but, of conceptual necessity, only in reference to something that A does.[54] A's performing an act in reference to which a desert-assessment of A coherently can be made is a conceptually necessary condition for A's desert. It is this feature of retributivism that suggests a natural type of justification of the VAR: (Of course the VAR is justified: (a) legal punishment is justified by desert, and (b) deserving punishment is incoherent except in light of something that the criminal defendant did.)

Any complete justification of the VAR must address not only the necessity of an act but also the voluntariness of that act. What is the word "voluntary" doing in MPC § 2.01?[55] Why include it? What, if anything, does it add? As mentioned previously, the ALI thinks that the term "voluntary" serves at least to emphasize an agent's control over her behavior.[56] And if a theory's main justificatory notion for legal punishment is a criminal defendant's desert, then, under that theory, the most natural reason to think that control is important is that it makes possible a defendant's moral responsibility for her conduct. For a defendant to deserve legal punishment for what she did, she must be morally responsible for what she did. And moral responsibility requires sufficient control.

Although desert requires moral responsibility, which in turn requires sufficient control, some retributivists argue that a defendant's desert can turn in part on factors over which the defendant lacks control. In addition to holding that a voluntary

53. MODEL PENAL CODE § 2.01 cmt. 1 (Official Draft and Explanatory Notes 1985) ("It is fundamental that a civilized society does not punish for thoughts alone."); MODEL PENAL CODE § 2.01 explanatory note on subsection (1) ("a fundamental predicate for all criminal liability [is] that the guilt of the defendant be based upon conduct, and that the conduct include a voluntary act or an omission to perform an act of which the defendant was physically capable. . . . [L]iability cannot be based upon mere thoughts, upon physical conditions, or upon involuntary movements.").

54. Here, I am ignoring the act/omission distinction and the point that an omission can (under a broadly retributivist scheme) be the basis of an agent's deserving punishment.

55. *See* MODEL PENAL CODE § 2.01 (1962).

56. *See supra* Part I.

act is necessary for desert, some[57] retributivists argue that when an agent voluntarily acts, that agent's desert can be conceptualized as a function of two elements—culpability and wrongdoing.[58] Culpability is a function of the *mens rea*, purpose, knowledge, recklessness, negligence, that accompanies the act.[59] Wrongdoing is a function of the badness of the results of the act.

For example, Abbott and Costello both recklessly drive automobiles in a busy part of town. By happenstance, Abbott manages not to hit anyone, barely missing Lewis, a pedestrian. But Costello hits a pedestrian, Clark. Abbott is less deserving of punishment than Costello,[60] even though both are equally culpable, for Abbott committed less wrongdoing than Costello. Another example: Abbott negligently drives a car and unintentionally hits and kills Lewis, a pedestrian. Costello intentionally sets out to kill a pedestrian, Clark, and purposefully drives right at Clark, hitting and killing Clark.

57. Some retributivists focus only on culpability and argue that wrongdoing has no independent moral relevance to an agent's desert. For example:

> I propose to consider what to make of a doctrine of the criminal law that seems to me not rationally supportable . . . . This is the doctrine—the harm doctrine, I'll call it—that reduces punishment for intentional wrongdoers (and often precludes punishment for negligent and reckless wrongdoers) if by chance the harm they intended or risked does not occur.

Sanford H. Kadish, *Supreme Court Review: Foreword*: *The Criminal Law and the Luck of the Draw*, 84 J. CRIM. L. & CRIMINOLOGY 679, 679 (1994).

58. Note that *wrongdoing* here is being used in a technical sense. The term does not refer simply to the performance of a wrongful action. As will be elaborated, *wrongdoing* refers to the badness of the results of what the defendant does—worse outcomes mean greater wrongdoing. *See* MOORE, PLACING BLAME, *supra* note 32, at 191 ("both culpability and wrongdoing matter to one's just deserts . . . . [T]here are two independent desert-bases, wrongdoing and culpability. . . . [T]o ask what punishment someone deserves is to ask how much wrong they did, and with what culpability they did the wrong.").

59. MODEL PENAL CODE § 2.02(1) ("Minimum Requirements of Culpability . . . . [A] person is not guilty of an offense unless he acted purposely, knowingly, recklessly or negligently, as the law may require, with respect to each material element of the offense.").

60. Here I leave open whether *less deserving of punishment* refers to whether Abbott is to be punished at all or to whether Abbott is to be punished less severely than Costello.

Abbott is less deserving of punishment than Costello, even though both have committed the same amount of wrongdoing, for Abbott is less culpable than Costello.

Michael Moore has elaborated and defended the idea that culpability and wrongdoing are independent desert-bases.[61] And he has clarified how the relationship between desert and culpability differs from the relationship between desert and wrongdoing. Assuming a voluntary act, culpability is both necessary and sufficient for desert. If an agent acts culpably, then the agent deserves some legal punishment. If an agent does not act culpably, then the agent does not deserve any legal punishment, even if the consequences of the act are really bad. On the other hand, wrongdoing is neither necessary nor sufficient for desert. For example, someone might deserve legal punishment for committing an inchoate crime that did not, because it was inchoate, generate any particularly harmful results. Or someone voluntarily but non-culpably might do something very harmful. In the latter case, a well-constituted agent would likely feel regret, but she would not deserve legal punishment. However, if an agent is culpable and, therefore, deserving of some punishment, then the amount of punishment deserved is in part a function of wrongdoing. Wrongdoing takes on independent significance, but only in the presence of culpability, whereas culpability always has independent significance on its own.

To summarize: According to retributivism, the justification of punishment turns on the defendant's desert. A defendant can have desert only if she performs an act over which she has the right sort of control to make her morally responsible for that act. In addition, she must be culpable in performing the act—that is, she must have done it purposely, knowingly, recklessly, or negligently.[62] Finally, assuming that she is culpable, her degree

61. *See* MOORE, PLACING BLAME, *supra* note 32, at 191–93.

62. Someone might object that strict liability crimes have no *mens rea* (culpability) requirement and that therefore a defendant who committed a strict liability crime could not deserve punishment for having committed it. However:

> There are two ways to construe strict liability crimes: (1) the crime has no *mens rea* requirements; or (2) the crime has *mens rea* requirements but any mental state on the part of

of desert can be affected by the badness of the results of her conduct.

III. The Problem of Moral Luck, the Problem of Free Will, and the Principle of Alternate Possibilities

As explained, according to the ALI, the VAR requires a voluntary act to assure that the criminal defendant has the requisite sort or amount of control over her conduct to legitimize imposing criminal responsibility upon her for that conduct. And central to retributivist justifications is the idea that this control must be sufficient to ground moral responsibility. To identify and elaborate the sort of control necessary to ground moral responsibility, it is helpful to consider two related problems—the problem of moral luck and the problem of free will.

A. *The Problem of Moral Luck*

Luck and control are tightly related. Luck can come into play when control is absent. Games of chance, involving random events not within anyone's control, require luck for success. Games like chess also can involve luck when something is not within a player's control. (I'm lucky that my opponent did not notice that devastating move during her turn at that point in the game. If she had, then I would have been checkmated within three moves. Instead, I went on to win.) That luck and lacking control are connected should not be confused with the stronger, arguably false, claim that luck is present in all cases in which there is a lack of control. For example, "[a]n event such as the rising of the sun this morning was entirely out of one's control, yet it is not at all clear that one is lucky the sun rose this

> the defendant meets them. To conceive of strict liability in the first way is to see strict liability crimes as involving a major departure from fundamental axioms of criminal law, particularly the principle according to which acts are never worthy of punishment in the absence of accompaniment by culpable mental states.

Yaffe, *supra* note 2, at 188–89. Conceptualizing strict liability crimes in the second way enables a response to the objection by respecting the fundamental axiom. *Id.*

morning, although it is surely a good thing that it did."[63] So if A has enough of the right kind of control over a situation, then how things turn out vis-à-vis A is not a matter of luck. But also, sometimes how things turn out vis-à-vis A is not a matter of luck even when A lacks control over a situation.

Additionally, it seems extremely plausible that how morally to assess A should depend only on factors that are, at least in some manner and to no small degree, under A's control. Whether A is morally responsible for a choice depends, at least in large part, on whether things are *really up to* A when A acts or chooses. Luck is not supposed to have anything to do with it.[64] To the extent that the results of what A does are not up to A, those results are not relevant to whether a given moral assessment of A[65] is correct.[66] This idea generates the problem

63. Andrew Latus, *Moral Luck*, INTERNET ENCYCLOPEDIA OF PHILOSOPHY (last visited March 19, 2016), http://www.iep.utm.edu/moralluc/.

64. Feinberg, *Problematic Responsibilities in Law and Morals*, *in* DOING & DESERVING: ESSAYS IN THE THEORY OF RESPONSIBILITY 32 (Princeton Univ. Press 1970) ("Moral responsibility . . . must be something one can neither escape by good luck nor tumble into through bad luck.").

65. This should not be confused with the view that the results of A's choices, even when not entirely up to A, are irrelevant to the moral assessment of A's choices (as permissible, morally required, morally worse than some other choices, and so forth). The problem of moral luck is a problem that arises when assessing agents, not their choices.

Also, this idea threatens the notion that wrongdoing is relevant to desert. The problem of moral luck motivates some retributivists to conceptualize desert as a function of voluntariness and culpability only. Such retributivists might argue, for example, that attempted murder should be punished as severely as murder when luck intervenes to render the attempted murder inchoate and the *mens rea* is the same in both cases.

66. *See* IMMANUEL KANT, GROUNDWORK OF THE METAPHYSICS OF MORALS 8 (Mary Gregor ed. & trans., Cambridge Univ. Press 1784).

> A good will is not good because of what it effects or accomplishes, because of its fitness to attain some proposed end, but only because of its volition, that is, it is good in itself . . . . Even if, by a special disfavor of fortune or by the niggardly provision of a step motherly nature, this will should wholly lack the capacity to carry out its purpose—if with its greatest efforts it should yet achieve nothing and only the good will were left (not, of course, as a mere wish but as the summoning of all means insofar as they are in our control)—then, like a jewel, it would still shine by itself, as something that has its full worth in itself. Usefulness or fruitlessness can neither add anything to this worth nor take

of moral luck. Moral luck comes into play "[w]here a significant aspect of what someone does depends on factors beyond his control, yet we continue to treat him in that respect as an object of moral judgment."[67] The problem posed by moral luck arises because: (a) at least in many cases, there seems to be such a thing as moral luck; and (b) both of the following are plausible[68] (where "(CP)" refers to the "Control Principle" and "(ML)" refers to "Moral Luck"):

> (CP) We are morally assessable only to the extent that what we are assessed for depends on factors under our control.
>
> (ML) [M]oral luck occurs when an agent can be *correctly* treated as an object of moral judgment, despite the fact that a significant aspect of what he is assessed for depends on factors beyond his control.[69]

An example should suffice to show why the idea that there can be such a thing as moral luck, as characterized by (ML), seems plausible to many. Abbott shoots at Lewis, and Costello shoots at Clark. Abbott's shot is not lethal because, out of nowhere, Horatio intercedes to take the bullet, striking Horatio's arm, that Abbott shoots at Lewis. Costello's shot is unimpeded and strikes Clark's head, instantly killing Clark. Even though it is beyond Abbott's and Costello's control whether there is someone like Horatio lurking around ready heroically to leap out of nowhere to shield potential shooting victims, what Abbott does is not as bad as what Costello does in the sense that what

anything away from it.

Id. at 8.

67. Thomas Nagel, *Moral Luck*, *in* MORTAL QUESTIONS 24, 26 (Cambridge Univ. Press 1979).

68. (CP) is plausible in the straightforward sense that, at least on the surface, it seems true. (ML) is plausible as a definitional statement, because it seems accurately to articulate the concept that we have in mind when we speak of moral luck.

69. Dana K. Nelkin, *Moral Luck*, *in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY (rev. ed. 2013), http://plato.stanford.edu/entries/moral-luck/.

Abbott is morally responsible for is not as bad as what Costello is morally responsible for. Costello killed someone, but Abbott did not. Costello is more blameworthy than Abbott. Although both (CP) and (ML) are plausible, if there are cases of moral luck, then in such cases we seem faced with the contradiction that things beyond someone's control both are and are not appropriate bases upon which to judge her morally.

How does the problem of moral luck help identify the sort of control that a criminal defendant must have over her conduct for her to act *voluntarily* and, thereby, be morally responsible for her conduct? The problem of moral luck seems to pose a problem primarily for what has been termed the harm doctrine, which "reduces punishment for intentional wrongdoers (and often precludes punishment for negligent and reckless wrongdoers) if by chance the harm they intended or risked does not occur."[70] That is, the most obvious target of the problem of moral luck seems to be the notion that wrongdoing can influence desert. But other implications of the problem of moral luck come into view once we consider that wrongdoing is not the only element of desert over which a defendant might lack control.

To see this more clearly, note that moral luck comes in at least four different varieties. Resultant luck, illustrated above in the head-shooting example, is "luck in the way things turn out."[71] Circumstantial luck is "luck in one's circumstances . . . .

70. Kadish, *supra* note 57, at 679.

71. Nelkin, *supra* note 69. Consider another example of resultant luck, involving a truck driver running over a child:

> The driver, if he is entirely without fault, will feel terrible about his role in the event, but will not have to reproach himself. Therefore this example of agent-regret is not yet a case of moral bad luck. However, if the driver was guilty of even a minor degree of negligence . . . then if that negligence contributes to the death of the child, he will not merely feel terrible. He will blame himself for the death. And what makes this an example of moral luck is that he would have to blame himself only slightly for the negligence itself if no situation arose which required him to brake suddenly and violently to avoid hitting a child. Yet the negligence is the same in both cases, and the driver has no control over whether a child will run into his path.

Nagel, *supra* note 67, at 28–29.

The things we are called upon to do, the moral tests we face, are importantly determined by factors beyond our control."[72] Constitutive luck concerns "the kind of person you are, where this is not just a question of what you deliberately do, but of your inclinations, capacities, and temperament."[73] And causal luck is "luck in how one is determined by antecedent circumstances."[74]

Causal luck suggests that the amount of harm that a defendant actually causes is not the only element of desert over which a defendant might lack control. As discussed, the harm actually caused is affected by factors *external* to the agent that the agent does not control. It is for this reason that it seems questionable to many to consider *wrongdoing*, in the technical sense of the term, a variable that can affect an agent's moral responsibility, and therefore her desert. But if we limit the factors relevant to assessing an agent's moral responsibility to those *internal* to the agent, then we escape the problem that arises from treating wrongdoing as a desert-basis. We can "admit that moral responsibility for external harm makes no sense and argue that moral responsibility is . . . restricted to the inner world of the mind . . . for here is a domain where things happen without the consent of uncooperative nature."[75] The agent, after all, is in control of her own mind.

But this strategy takes us only so far. Joel Feinberg suggests an example involving two virtually identical aggressors—Hotspur and Witwood. Each aggressor is imagined,

72. Nagel, *supra* note 67, at 33.

73. *Id.* at 28. To elaborate:

> Since our genes, care-givers, peers, and other environmental influences all contribute to making us who we are (and since we have no control over these) it seems that who we are is at least largely a matter of luck. Since how we act is partly a function of who we are, the existence of constitutive luck entails that what actions we perform depends on luck, too.

Nelkin, *supra* note 69.

74. Nagel, *supra* note 67, at 28. Arguably, the category of causal luck is superfluous "because circumstantial and constitutive luck seem to cover the same territory. Constitutive luck covers what we are, while circumstantial luck covers what happens to us. Nothing else seems to remain that can play a role in determining what we do." Latus, *supra* note 63.

75. Feinberg, *Justice and Personal Desert*, *supra* note 50, at 33.

in separate incidents, to slap a victim called Hemo in the face. Hemo turns out to be a hemophiliac. When Hotspur slaps Hemo he cuts Hemo's mouth, and Hemo bleeds to death. When Witwood strikes Hemo something external, that is, not involving Witwood's mind, happens that prevents Hemo's death. Comparatively speaking, Hotspur is morally unlucky because he is responsible for greater wrongdoing—Hemo's death instead of Hemo's nonfatal injury—than Witwood. Witwood enjoys morally better fortune. Feinberg asks us to imagine "rewinding" the episode in each case to a point before the aggressor slaps Hemo:

> The same good fortune is possible at earlier "internal" stages. For example, at the stage when Hotspur would begin to burn with rage, a speck of dust throws Witwood into a sneezing fit, preventing any rage from arising. He can no more be responsible for a feeling he did not have than for a death that did not happen. Similarly, at the point when Hotspur would be right on the verge of forming his intention, Witwood is distracted at just that instant by a loud noise. By the time the noise subsides, Witwood's blood has cooled, and he forms no intention to slap Hemo . . . . [I]n whatever sense legal responsibility for external states can be contingent on factors beyond one's control and therefore a matter of luck, in precisely the same sense can "moral" responsibility for inner states also be contingent and a matter of luck.[76]

To put this idea another way, causal luck can affect what happens inside the head, just as resultant luck can affect what happens outside the head. The problem of moral luck therefore points toward a sort of actor's control that seems necessary to ground the actor's moral responsibility. That type of control needs to overcome how causal luck affects what goes on in the actor's head when she voluntary acts.

76. *Id.* at 35.

B. *The Problem of Free Will*

Causal luck also plays a role in a similar problem concerning the relationship between free will and determinism, sometimes called the problem of free will. The rough idea behind the problem of free will is: If everything we do is causally determined, then it is not really up to us what we do—we are not really free because we are not really in control of what we do. On the other hand, if what we do is not causally determined by anything, then, again, we are not really free because we are not really in control of what we do. If nothing determines what we do, then neither do we. So although, intuitively, it seems as if it is really up to us what we do, at least sometimes, the foregoing seems to rule this out as a possibility—hence, the problem.

It is important to formulate the problem of free will precisely to decide whether it admits of a possible resolution and to understand what differences there may be among competing solutions. A set of standard terms has been developed to help accomplish this. I shall adopt the definitions offered by one expert, Peter van Inwagen:[77]

> "*Determinism* is the thesis that the past and the laws of nature together determine, at every moment, a unique future . . . ."[78]
>
> *Indeterminism* is "[t]he denial of determinism . . . ."[79]
>
> The *Free-Will Thesis* is the thesis that "we are sometimes in the following position with respect to a contemplated future act: we simultaneously have both the following abilities: the ability to perform that act and the ability to refrain from performing that act (This entails that we *have been* in the following position: for something we did do, we were at some point prior to our doing it

77. *See generally* van Inwagen, *How to Think*, *supra* note 9.
78. *Id.* at 330 (emphasis added).
79. *Id.* (emphasis added).

> able to refrain from doing it, able not to do it)."[80]
>
> "*Compatibilism* is the thesis that determinism and the free-will thesis could both be true. . . ."[81]
>
> "*[I]ncompatibilism* is the denial of compatibilism."[82]
>
> "*Libertarianism* is the conjunction of the free-will thesis and incompatibilism (Libertarianism thus entails indeterminism)."[83]
>
> "*Soft determinism* is the conjunction of determinism and the free-will thesis (Soft determinism thus entails compatibilism)."[84]

With these definitions at hand, we can formulate the problem of free will as follows: "Free will seems to be incompatible both with determinism and indeterminism. Free will seems, therefore, to be impossible. But free will also seems to exist. The impossible therefore seems to exist. A solution . . . would be a way to resolve this apparent contradiction."[85]

One possible solution would be to deny the existence of free will, to deny the free-will thesis. But if we deny the free-will thesis, then we run into trouble holding two other theses, both of which seem to be correct.[86] The first is that *ought implies can*. If A lacks the ability to do, or refrain from doing, something, then A could not be morally required to do, or refrain from doing, it.

80. *Id.* at 329 (emphasis added). Famously, Harry Frankfurt raised a potential problem concerning this way of formulating the free-will thesis if we understand the free-will thesis as necessary for moral responsibility. *See infra* Part II.C where I shall consider and adopt Frankfurt's point.

81. van Inwagen, *How to Think*, *supra* note 9, at 330 (emphasis added).

82. *Id.* (emphasis added).

83. *Id.* (emphasis added).

84. *Id.* (emphasis added).

85. Peter van Inwagen, *Free Will Remains a Mystery: The Eighth Philosophical Perspectives Lecture*, 14 PHIL. PERSP. 1, 11 (2000) [hereinafter van Inwagen, *Free Will Remains*].

86. Although some might dispute these two theses, this paper assumes that both are true.

For example, I could not morally be required to fly like Captain Marvel.

If Laurel and Hardy are two random strangers who happen to meet, then it seems obvious, in the absence of any extraordinary countervailing factors, that Laurel should not kill Hardy. If the free will thesis is false, however, then whenever Laurel is faced with a choice at a certain time and makes a particular decision, the decision that Laurel makes turns out to be the only one that Laurel had the ability to make at that time. But then, if Laurel kills Hardy, then Laurel never had the ability to refrain from killing Hardy. So if we deny the free will thesis while trying to hold onto the thesis that *ought implies can*, then we must give up the idea that if Laurel kills Hardy, then Laurel does something that Laurel ought not to have done. Further, if Laurel was not morally required to refrain from killing Hardy, then it becomes difficult to see how to justify: (a) having a system of legal punishment under which killers are punished, at least in part, because they are morally responsible for killing and (b) punishing Laurel under that system. For when Laurel killed Hardy, Laurel did not do anything morally impermissible.

The second, related thesis that is threatened by denying the free-will thesis is that A is sometimes morally responsible for what A does.[87] If every action[88] that A performs is the only one that A in fact ever has the ability to perform, then it seems implausible to ascribe moral responsibility to A for anything that A does. The threat that A is never morally responsible for what A does persists, even when we ignore the *external* aspects of what A does and focus only on what might be termed A's volition,

87. The notion of moral responsibility here is:

> [A]n absolute responsibility wholly within the power of the agent . . . . To be morally responsible . . . is not [in itself] to be liable to any kind of official action or even to unofficial informal responses such as acts of blaming. Moral responsibility . . . is liability to charges and credits on some ideal record, liability to credit or blame . . . . This record in turn can be used for any one of a variety of purposes—as a basis for self-punishment, remorse, or pride, for example.

Feinberg, *Justice and Personal Desert*, *supra* note 50, at 30–31.

88. Here, I am glossing over the difference between acting and omitting.

the *executory*[89] mental state that occurs in A's brain just before A acts. If every facet of A's *internal* psychological life is uniquely determined by past events, in principle, that occur before A is even born, in combination with the laws of nature, then causal luck affects even the formation of A's volitions.[90] Thus, the issue illustrated earlier by the example involving Hotspur, Witwood, and Hemo arises again, this time in connection with the problem of free will. If Hotspur forms the volition to strike Hemo, then that volition is the only volition that was ever within Hotspur's ability to form. So whether Hotspur forms the volition to strike Hemo is a matter of Hotspur's luck. When even Hotspur's volitions are in this sense not under Hotspur's control, it begins to seem impossible for Hotspur ever to be morally responsible for what he does. So solving the problem of free will by denying the free-will thesis seems off the table for anyone who thinks that we are ever morally responsible for what we do.[91]

Assuming that the free-will thesis is true and that we are at least sometimes morally responsible for what we do, there seems to be two ways to show how free will is possible—either convincingly argue that determinism is compatible with the free-will thesis or convincingly argue that indeterminism is. At first blush, it may seem that the best option is to argue that indeterminism is compatible with free will. Causal determination by factors entirely outside of one's control seems

89. *See infra* Part IV where I clarify the idea that a volition is a type of executory mental state, a type of intention.

90. Nagel, *supra* note 67, at 35.

> If one cannot be responsible for consequences of one's acts due to factors beyond one's control, or for antecedents of one's acts that are properties of temperament not subject to one's will, or for the circumstances that pose one's moral choices, then how can one be responsible even for the stripped-down acts of the will itself, if they are the product of antecedent circumstances outside of the will's control?

Id.

91. Since retributivists think that justly punished criminals must have been morally responsible for committing their crimes, I shall assume that retributivists accept the free-will thesis. It is not as clear that utilitarians must accept the free-will thesis, since utilitarian justifications of punishment do not primarily turn on whether the criminal defendant is morally responsible for what she does.

to pose the more serious threat to moral responsibility. If factors over which I lack any control determine everything that I do, or even think, feel, or will, then how can anything that I do ever really be up to me in the manner that the free-will thesis suggests?

But at least certain forms of indeterminism also seem incompatible with the free-will thesis. These include non-causal and event-causal varieties. According to non-causal indeterminism, "actions are free if the simple [mental] actions at their core are uncaused."[92] But if the cores of such actions are uncaused, then the agent does not cause them and, therefore, does not control them. The problem seems just as vexing when nothing has control as when something other than the agent does.[93]

According to event-causal indeterminism, certain agent-involving events cause those of an agent's actions for which the agent is morally responsible. When those agent-involving events non-deterministically cause a free action:

> [T]he agent exercises . . . a certain variety of active control (which is said to consist in the action's being caused . . . by those agent-involving events), the action is performed for a reason, and there remains, until she acts, a chance of the agent's not performing that action.[94]

92. Timothy O'Connor, *Why Agent Causation?*, 24 PHIL. TOPICS 143, 146 (1996) [hereinafter O'Connor, *Why Agent Causation?*].

93. ROBERT NOZICK, PHILOSOPHICAL EXPLANATIONS 292 (Harvard Univ. Press 1981) ("Random acts and caused acts alike seem to leave us not as the . . . originators of action but as an arena, a place where things happen, whether through earlier causes or spontaneously."). Along similar lines:

> An action's being non-determined . . . is not sufficient for it to be free . . . . If we acted in the way uranium 238 emits alpha particles, determinism would be false but (unless we are greatly mistaken about uranium 238) we would not thereby have free will.

Id. at 299.

94. Randolph Clarke & Justin Capes, *Incompatibilist (Nondeterministic) Theories of Free Will*, *in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY (rev. ed. 2013), http://plato.stanford.edu/entries/incompatibilism-theories/.

But such agent-involving events do not seem to provide the agent the requisite control any more than uncaused events do. "[T]he relevant causal conditions antecedent to a decision . . . would leave it open whether this decision will occur . . . . [A]nd whether [the decision occurs] is not settled by the agent. Hence, the agent lacks the control required for being morally responsible for the decision."[95]

So where does this leave us? It would seem that the retributivist must choose between libertarianism and soft determinism.[96] These two positions yield competing solutions to the problem of free will that might preserve moral responsibility, since according to both, the free-will thesis is true. Consider the following two propositions:

> (1) When the moment of choice arrives, A has the ability to do X and the ability not to do X.
>
> (2) Given the past, before A's birth, and the laws of nature, it is uniquely determined, even before A is born, that A does X when the moment of choice arrives.

If we understand the term *ability* in (1) to refer to the sort of control that an agent must have over her conduct to be morally responsible for it, then the libertarian thinks that (1) and (2) could not both be true. The libertarian's intuition is that because neither the past state of the world, before A's birth, nor the laws of nature are at all up to A, anything that the past and the laws of nature necessitate is not up to A either. A could not have the ability not to do something, here, X, that is necessitated that way.

The soft determinist, in contrast, sees no problem with (1) and (2) both being true. The soft determinist adopts the view that A has the ability *to do* or *not to do* X, provided that A *would*

95. Derk Pereboom, *Is Our Conception of Agent-Causation Coherent?*, 32 PHIL. TOPICS 275, 276 (2004).

96. By way of reminder: Libertarianism is the conjunction of the free-will thesis and incompatibilism. Soft determinism is the conjunction of the free-will thesis and determinism.

have done or *not have done* X had A chosen to or not to.[97] Once we understand A's ability to do X to amount to *A would have done X if A had chosen to*, the tension with determinism is eliminated. As the soft determinist might say: *So what if A was necessitated by factors completely beyond A's control not to choose to do X at the moment of choice? A had the ability to do X because A would have done X if A had chosen to.*

To justify the VAR in relation to a defendant's desert, the retributivist must deal with the problem of moral luck and the problem of free will in a way that preserves moral responsibility. To do that, the retributivist must adopt either libertarianism or soft determinism. This paper argues that the most plausible option for the retributivist is to adopt libertarianism in combination with a view called agent causalism, according to which we are the causes of our conduct when we voluntarily act and are morally responsible for what we do, as opposed to events that cause our conduct. *For retributivist justifications of the VAR to be plausible, agent causalism must be true.*

Before proceeding to that argument, it is important to reconsider an important aspect of how the problem of free will is formulated. The standard type of formulation outlined above presupposes what has been termed the Principle of Alternate Possibilities ("PAP"). The PAP, however, has been seriously called into question. So before continuing with the main line of argument, I shall explain, following a well-known observation by Harry Frankfurt,[98] why the PAP is false and consider whether this has any impact on the retributivist's choice between libertarianism and soft determinism.

C. *The Principle of Alternate Possibilities*

Recall the free-will thesis:

97. To put it a slightly different way, assume that A chose not to do X and was determined so to choose. The soft determinist holds that A had the ability to do X as long as: A would have done X if A had formed the volition to do X. Presumably, it would have been A's volition to do X that caused A's doing X had A formed that volition.

98. *See generally* Harry G. Frankfurt, *Alternate Possibilities and Moral Responsibility*, 66 J. PHIL. 829 (1969).

> [W]e are sometimes in the following position with respect to a contemplated future act: we simultaneously have both the following abilities: the ability to perform that act and the ability to refrain from performing that act (This entails that we *have been* in the following position: for something we did do, we were at some point prior to our doing it able to refrain from doing it, able not to do it.).[99]

If we understand free will to be important at least in part because it secures the possibility of moral responsibility, then formulating the free-will thesis this way seems implicitly to presuppose the PAP: "[A] person is morally responsible for what he has done only if he could have done otherwise."[100] Despite the PAP's initial plausibility, Harry Frankfurt provided a persuasive reason to think it false:

> [T]here may be circumstances that make it impossible for a person to avoid performing some action without those circumstances in any way bringing it about that he performs that action. It would surely be no good for the person to refer to circumstances of this sort in an effort to absolve himself of moral responsibility for performing the action in question. For those circumstances, by hypothesis, actually had nothing to do with his having done what he did.[101]

To illustrate his main idea, Frankfurt creates an example in which someone called Black is prepared to take steps to assure that another person, Jones, does what Black wants Jones to do. (Frankfurt emphasizes that the idea that it is a person—Black—that is doing this is irrelevant; a non-personal causal agency would do just as well.) What Black's steps are is left to the imagination of the reader, as long as the reader would

99. van Inwagen, *How to Think*, *supra* note 9, at 329.
100. Frankurt, *supra* note 98, at 829.
101. *Id.* at 837.

acknowledge that those steps assure that Jones can do only what Black wants Jones to do.[102]

Frankfurt's example:

> Suppose someone—Black, let us say—wants Jones[] to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones[] is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones[] is going to decide to do something *other* than what he wants him to do. If it does become clear that Jones[] is going to decide to do something else, Black takes effective steps to ensure that Jones[] decides to do, and that he does do, what [Black] wants him to do. Whatever Jones[]'s initial preferences and inclinations, then, Black will have his way . . . . Now suppose that Black never has to show his hand because Jones[], for reasons of his own, decides to perform and does perform the very action Black wants him to perform. In that case, it seems clear, Jones[] will bear precisely the same moral responsibility for what he does as he would have borne if Black had not been ready to take steps to ensure that he do it. It would be quite unreasonable to excuse Jones[] for his action, or to withhold the praise to which it would normally entitle him, on the basis of the

102. *See generally id.* at 835.

> What steps will Black take, if he believes he must take steps, in order to ensure that Jones[] decides and acts as [Black] wishes? Anyone with a theory concerning what "could have done otherwise" means may answer this question for himself by describing whatever measures he would regard as sufficient to guarantee that, in the relevant sense, Jones[] cannot do otherwise.

Id.

fact that he could not have done otherwise.[103]

Because he concludes that the PAP is false, Frankfurt proposes a similar, modified version that he thinks may be true. Call it the "PAP′": "[A] person is not morally responsible for what he has done if he did it only because he could not have done otherwise."[104] The force of the term *because* in the PAP′ is not justificatory, but explanatory. That is, if the correct causal explanation of how someone behaves on a certain occasion includes her inability to behave in any other way on that occasion, then she is not morally responsible for her behavior on that occasion.

As previously explained, the retributivist who wants to justify the VAR faces a choice between libertarianism and soft determinism. Facially, the PAP′ is consistent with either position. If causal determinism is true, then, according to the libertarian, no one is ever morally responsible for what she does because her conduct is always caused by factors, such as the history of the world and the deterministic laws of nature, that guarantee that she could not have done otherwise. It is largely for that very reason that the libertarian endorses indeterminism. The soft determinist would agree that those same factors, the history of the world and the laws of nature, always ultimately necessitate someone's behavior. But the soft determinist would insist that those factors are consistent with her ability, at the time of choice, to do otherwise. She could have done otherwise because she would have done otherwise if she had chosen to. In other words, the soft determinist can accept the PAP′ and determinism simultaneously, without concern that our status as morally responsible agents is in peril.

In light of the problem of free will, is there any reason for a retributivist who aspires to justify the VAR not to be a soft determinist? I submit that there is and that it is compelling. Soft determinism's view about when it is within someone's ability[105] to do something is extremely implausible. If every detail of A's psychology over the course of A's entire life is

103. *Id.* at 835–36.

104. *Id.* at 838.

105. Here, a person's ability to do something must encompass the sort of control necessary to ground moral responsibility for doing that thing.

necessitated by factors entirely outside of A's control, then it is extremely implausible that A has the ability to do X simply because A would have done X had A chosen to. Human persons like A have beliefs and desires that, at least sometimes, can rationally influence and explain how they behave. A tornado, in contrast, is not that sort of thing. And A's being the sort of thing that can make choices on the basis of reasons, by entering into various intentional[106] states that play a role in causing those choices, seems to be a *necessary* condition for A to be the sort of thing that properly can be subject to moral assessment. But being capable of choosing for reasons by entering into intentional states is not *sufficient* for moral responsibility. If the existence, content, and causal effects of A's mental states are entirely outside of A's control—if they are not at all up to A—then it is hard to see how those mental states, or counterfactual claims about choices resulting from them, could explain how A could be more of a morally responsible agent than a tornado could ever be. And it would be absurd to claim that a tornado might deserve legal punishment for its *conduct*.

One of the statements in the formulation of the problem of moral luck considered in Part III.A was:

> (CP) We are morally assessable only to the extent that what we are assessed for depends on factors under our control.

The soft determinist's conception of control cannot justify the sort of moral assessment to which (CP) refers.[107]

106. Here, the noun form of *intentional* is the technical term *intentionality*. *Intentionality* refers to "that property of many mental states and events by which they are directed at or about or of objects and states of affairs in the world." JOHN R. SEARLE, *The Nature of Intentional States, in* INTENTIONALITY: AN ESSAY IN THE PHILOSOPHY OF MIND 1 (Cambridge Univ. Press 1983). Intentionality here means aboutness, not purposiveness. "[I]ntending to do something is just one form of [i]ntentionality along with belief, hope, fear, desire, and lots of others." *Id.* at 3.

107. Beyond what I have written in the preceding paragraph, I have nothing worthwhile to offer the soft determinist that might convince her that her view is incorrect. And I recognize that there are many capable, convinced soft determinists. Soft determinists need to explain how their conception of an agent's control makes defensible the moral assessment of the agent. And there is an enormous literature that tries to do that. Addressing those arguments

IV. Agent Causalism

In light of the problem of free will, therefore, it seems that libertarianism is the way to go for retributivists seeking to justify the VAR. As explained earlier, however, the idea that some events are uncaused or that some events non-deterministically cause free choices does not help make room for moral responsibility or desert. This is where agent causalism comes in.

Whenever A voluntarily does something, we can ask whether A, the agent, is ever the cause of what A does. Alternatively, we can ask whether it is always some event or chain of, perhaps neurophysiological, events—distinct from A—that causes what A does. Those who hold that only events ever stand in direct causal relations are called event-causalists;[108] and those who think that sometimes agents stand in direct causal relations to events are called agent-causalists. Event-causalists believe that all causes are events and that, correspondingly, there is only one kind of basic causal relation—a relation whose subject and object are both events. "[T]he event-causalist [contends] that the causation of events intrinsic to . . . actions by the intendings [i.e., volitions] of an agent is just a matter of 'ordinary' event-causation."[109] An agent-causalist, on the other hand, believes that voluntary actions involve an irreducible causal relation whose subject is the agent herself. According to agent-causalism, some causes are substances,[110]

exceeds what I can do in this paper.

108. *See* John Bishop, *Agent-causation*, 92 MIND 61 (1983).

> Does every intentional action involve an irreducible causal relation whose subject is, not an event or sequence of events, but the agent himself? Those who say not [can be called] *event-causalists* . . . . To them, the causal component in intentional action is a matter of 'ordinary' causal relations amongst events, and the explanation of behaviour as intentional action is just a species of 'ordinary' causal explanation.

Id. at 61.

109. *Id.* at 63.

110. The term "substances" is a metaphysical term. But although it is

called agents, and there is a corresponding distinctive and irreducible form of agent-causal relation in addition to the type of causal relation that can hold between events. This irreducible[111] agent-causal relation constitutes an agent's control over her actions. And this sort of control makes an agent morally responsible for her choices because it is a distinctive sort of power: "[A] causal power, fundamentally as a substance, to cause a decision without being causally determined to do so."[112]

Timothy O'Connor offers further elaboration:

> Wherever the agent-causal relation obtains, the agent bears a *property* or set of properties that is 'choice-enabling' (i.e., in virtue of such properties, the agent has a type of causal power which . . . we may term "active power"). But this 'active power'—the causal power in virtue of which one has freedom of will—is not characterized by any function from circumstances to effects (as is the case with event causal powers). For the properties that confer such a capacity do not themselves . . . necessitate or make probable a certain effect. Rather, they . . . *make possible* the direct, purposive bringing about of an effect *by the agent* who bears them. Such properties thus play a different functional role in the associated causal process . . . . [T]hese properties give rise to a fundamentally different type of causal power—

abstract, it should not be considered unclear, abstruse, or highfalutin. "There is an ordinary concept in play when philosophers discuss 'substance', and this . . . is the concept of *object*, or *thing* when this is contrasted with properties or events." Howard Robinson, *Substance*, *in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY (rev. ed. 2014), http://plato.stanford.edu/entries/substance/.

111. "Irreducible" refers to the impossibility of reducing an agent-causal relation in terms of event-causal relations alone. In other words, agent-causation is not ultimately a complicated form of event-causation in disguise. Agent-causation is a fundamental sort of causation, just as the fundamental forces of physics are not reducible to one unifying force. (At least, so far as I know, the fundamental forces of physics have not been unified yet—maybe they will be, and maybe they will not. If we discover that the fundamental forces are unified, we would realize that they were not really fundamental in the way we thought they were.)

112. Pereboom, *supra* note 95, at 278.

> one that in suitable circumstances is exercised at will by the agent, rather than of necessity, as with objects that are not partly self-determining agents.[113]

This power must also be reasons-responsive—the agent can act rationally, for reasons, when she exercises this power. Reasons must be capable of guiding, explaining, and justifying what an agent does when she freely acts. In other words, when an agent voluntarily acts, the agent-causal power operates in concert with her intentional states, beliefs, desires, etc., such that the contents of those states play an indispensable role in explaining what the agent does. Acting with moral responsibility requires the capacity to act on the basis of practical reason.[114]

Obviously missing from the explanation of agent causalism offered thus far is a detailed specification of the set of properties and circumstances such that, if a substance has those properties in those circumstances, then the substance is an agent that can directly bring about an event in response to reasons to do so. But there is not anything incoherent or implausible about there being such a specification, if only we could identify it. A pane of

113. O'Connor, *Why Agent Causation?*, *supra* note 92, at 145. Along similar lines:

> [T]he agency theory . . . affirms the completely general claim . . . that objects have causal powers in virtue of their properties, so that objects sharing the same properties share the same causal capacities . . . . [S]ome properties contribute to the causal powers of the objects that bear them in a very different way from the event-causal paradigm, in which an object's possession of property P in circumstance C necessitates or makes probable a certain effect. On this alternative picture, a property of the right sort can (in conjunction with appropriate circumstances) make possible the direct, purposive bringing about of an effect by the agent who bears it.

Timothy O'Connor, *Agent Causation*, *in* AGENTS, CAUSES, AND EVENTS: ESSAYS ON INDETERMINISM AND FREE WILL 173, 177 (Timothy O'Connor ed., Oxford Univ. Press 1995) [hereinafter O'Connor, *Agent Causation*].

114. *See infra* Part V where I shall elaborate how agent-causal power might fit within a picture of how an agent acts for reasons by considering the anatomy of a voluntary act.

glass is fragile, and in certain circumstances it will either certainly shatter or have a high probability of shattering. Presumably, there is a detailed technical specification of the micro-structural properties of the pane of glass that, if only we could identify it, would illuminate why the pane of glass behaves that way in those circumstances.[115] What rules out a specification of a thing's properties and a set of circumstances that illuminates why that thing, and any other thing that has those properties in those types of circumstances, is capable of being an agent that can directly cause an event in response to reasons? Of course, we eventually might learn enough, empirically, about how our brains work that we become convinced that there is no accurate specification of our properties in any set of circumstances that could explain how we could be agents.[116] If that were to happen, then retributivist approaches to justifying the VAR would be in serious trouble.

How does agent-causation empower people to be morally responsible for what they do? When an actor freely acts, she *agent-causes*[117] her decision to do so. Actors are morally responsible for what they do when they agent-cause their decisions to act. A Frankfurt-style scenario suggested by William Rowe can help clarify this idea. Imagine that Jones is deciding whether (a) to keep some money that does not belong to him or (b) to return the money to its rightful owner, who needs

115. I suspect that we already have such a specification available, based on our empirically-acquired knowledge of glass.

116. For example, one empirical study seems to indicate that certain voluntary decisions are caused by brain activity that occurs before the subject is conscious of making a decision. This suggests that it is not the subject herself who causes the voluntary conduct—preconscious brain events and brain states cause it instead. *See* Benjamin Libet et al., *Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential): The Unconscious Initiation of a Freely Voluntary Act*, 106 BRAIN 623 (1983). Although I do not believe that this study, standing alone, is a serious threat to agent causalism, if enough sophisticated studies of complex human decision making were performed that generated results inconsistent with the agent-causalist picture, then agent causalism would be falsified, as least as to us. It would turn out that even if there were agents, we would not be included among them.

117. To say that an actor *agent-causes* an event is to mean that the above-mentioned irreducible causal relation (whose subject is the actor herself) is realized and that the event occurs in virtue of that causal relation being realized.

that money more than Jones does. Jones thinks that the morally right thing to do is to return the money. But he selfishly decides to keep it for himself under the following conditions:

> No outside influence or internal desire or want caused him to decide to keep the money. He was free to cause and free not to cause his decision to keep the money. As it happened, he followed his selfish desire, rather than the advice of his conscience, and [agent-caused] his decision to keep the money, having it within his power, nevertheless, not to have [agent-caused] that decision. However, had he been about to [agent-cause] the decision to return the money, the devil, let us suppose, would have directly caused in him the decision to keep the money, effectively preventing any decision or action on his part to return the money . . . . In a way, given the steady resolve of the devil, it is up to our agent whether he himself *or* the devil will be responsible for his decision to keep the money. By exercising his power to [agent-cause] his decision to keep the money, he makes himself responsible for that decision. Had he not [agent-caused] that decision, the devil, and not he, would have been responsible for his decision to keep the money. And had he not [agent-caused] his decision to keep the money, then, at long last, we would have a case in which someone might *truthfully* say: "the devil made me do it."[118]

In this example, when the devil does not make Jones keep the money, Jones agent-causes his decision to keep it. And when the devil makes Jones do it, Jones does not agent-cause his decision—the devil causes it by causing in Jones a necessitating volition to keep the money, which in turn causes Jones to keep it. Only when the devil does not make him do it is Jones morally responsible for keeping the money. And it is up to Jones whether he is morally responsible, since Jones is in control of whether or

118. William L. Rowe, *Free Will, Moral Responsibility, and the Problem of "Oomph,"* 10 J. ETHICS 295, 299 (2006).

not he agent-causes the decision. If it were a crime for Jones to keep the money, then, in principle, Jones might deserve legal punishment for keeping the money if the devil did not make him do it. And all of this is true even though Jones could not have done otherwise than to keep the money.

As stated previously, this paper's main thesis is that *for retributivist justifications of the VAR to be plausible, agent causalism must be true*. This thesis is significant because agent causalism is contentious and may be false. And if agent causalism is false, then retributivism could not play any role in substantiating the VAR, the fundamental legal precondition of ever imposing criminal liability upon anyone. To be contentious, agent causalism must, to some degree, be plausible. If agent causalism were entirely implausible, no one would take it seriously as a possibility. And to be at all plausible, agent causalism must at least be coherent. Although my objective is not thoroughly to defend agent causalism, I must explain how agent causalism is coherent and has at least some plausibility. Otherwise, this paper's main thesis would not be significant in the way that I claim.[119]

Why might someone think that agent causalism is incoherent or extremely implausible? Before considering what I take to be two of the more challenging objections to agent causalism, let me quickly address what I consider to be, at most, a couple of superficial reasons to dismiss agent causalism. A critic might believe that agent causalism is committed to non-physical substances, souls, say, or to supernatural phenomena and that such commitments are farfetched. Whether or not such things are farfetched, agent causalism is not committed to them. An agent causalist does think that agents are substances (things other than properties or events) and that this is important. An agent causalist also thinks: (a) that these substances are able to cause events without being determined to do so; (b) that these

119. If agent causalism were incoherent or entirely implausible, then this paper's main thesis would still be significant. For if it were obvious that agent causalism was incoherent or for some other reason was not true, then it would follow from this paper's thesis that there would be no plausible way for retributivist justifications of the VAR to work. But I think that there may be a plausible way for retributivist justifications of the VAR to work. At least, I am not prepared to rule out such a possibility. So I must explain how agent causalism is coherent and is at least somewhat plausible.

substances can thereby be morally responsible for their conduct; and (c) that (a) and (b) are important. But there is no reason to think that according to agent causalism agents are non-physical or that agent-causation is not a natural type of causal relation.[120]

For example, an agent causalist might think that organisms that can agent-cause events came to exist through evolutionary processes driven largely by natural selection, just as organisms that can fly, such as birds, came to exist through evolutionary processes driven largely by natural selection. There is no more reason to think that agents are non-physical or that the causal relations involved in an agent's ability to agent-cause events are supernatural than there is to think that birds are non-physical or that the causal relations involved in a bird's ability to fly are supernatural.

There are, however, more engaging objections to agent causalism. One such objection concludes that either agent causalism is false or, even if true, agent causalism can account for an agent's control over her conduct no better than the view that free actions are triggered by uncaused events can. Recall that uncaused events seem unable to make moral responsibility possible because an event under nothing's control is not under an actor's control. So if the actor's, allegedly, free actions are caused by uncaused events, then her, allegedly, free actions are not under her control, and she therefore could not be morally responsible for them.

The objection can be posed as follows: Every episode of agent causation features an agent, S, bringing about an event, *e*. So every episode of agent causation is itself a complex event with the structure: S's causing *e*. Call that complex structured event "E." What causes E? There seem to be only two available answers—either some earlier event causes E or S does. If an earlier event causes E, then we face anew the same problem that agent causalism was supposed to help us solve. Either that previous event was uncaused, or caused—maybe remotely—by an uncaused event, or determinism is true. But if S causes E, then an infinite regress ensues. For S to be morally responsible for conduct on a certain occasion, S must agent-cause an infinite

120. An agent causalist might think such things, but agent causalism does not imply them.

number of events on that occasion: *e*; E; (S's causing E); (S's causing (S's causing E)); (S's causing (S's causing (S's causing E))); and so on. And it is absurd to think that every time S freely acts S literally performs an infinite number of acts. So if S causes E, then agent causalism must be false.

The agent causalist can respond, however. One possibility is to accept the infinite regress but insist that it is not vicious. For example, imagine that you walk in a straight line from the center of a room to one of the room's walls, touching the wall.[121] Call that action—your walking to the wall—"*a*". Before you complete *a*, you complete another action—walking halfway to the wall. After you walk halfway but before you get all the way to the wall, you complete a third action—walking three-quarters of the way to the wall—and so on. In performing a quotidian action such as *a*, you perform an infinite number of actions. But this infinite regress is not vicious. If it were, then perhaps Zeno of Elea would have succeeded in showing that, appearances notwithstanding, you never get from point A to point B.[122] ("Sure, every time I freely act I perform an infinite number of actions. But so what? Every time I move from one place to another I perform an infinite number of actions. Where's the problem?")

Another possible response to the *infinite regress* objection has been suggested by Timothy O'Connor. Call events such as E[123] *causally complex events*. Run-of-the-mill events that are not themselves instantiations of causal relations, call them *causally simple events*, lack the internal structure of causally complex events. There are at least two types of causally complex events—event-causal events and agent-causal events. Event-causal events have the following structure: (E_1's causing E_2), where E_1 and E_2 are events (themselves either causally complex or causally simple). Agent-causal events have the following structure: (S's causing E), where S is an agent and E is an event (itself either causally complex or causally simple):

121. Assume for the sake of this example that space-time is continuous.

122. "That which is in locomotion must arrive at the half-way stage before it arrives at the goal." ARISTOTLE, PHYSICS VI: 9, 239b10 (recounting one of Zeno's paradoxes of motion).

123. Recall that E was (S's causing *e*). As will very shortly be explained, E is an example of an *agent-causal event*.

> [I]nstantiations of . . . (causally complex events) are not themselves directly on the receiving end of other causal relations—instead . . . (causally simple . . . events) are. Causings are the *producings* of events, rather than what are *produced* (in the first instance). Compare an ordinary case of an event-causal process (consisting of event *F*'s causing event *G*) being caused by some further *event E*. Surely this can consist only in *E*'s causing *F*, the front-end relatum of the complex event . . . . If this is right, then an *agent*-causal event could not be caused for the simple reason that the cause in this case is not an event.[124]

To unpack and amplify O'Connor's response: The objection's infinite regress gets going because it is assumed that any causally complex event, understood as a whole, must be brought about by something else, either another event or an agent. But this assumption is false.

> (1) – In the case of an event-causal event (F's causing G), nothing causes (F's causing G) as a whole—causal relations themselves never get caused like that. Instead, an event E or an agent A causes event F, the first component of the complex whole (F's causing G). And if F is caused in either of those ways, everything is accounted for—in particular, there are no uncaused events leftover. For example, if E causes F and F causes G, then we have as much of an explanation of (F's causing G) as there is to have.
>
> (2) – In the case of an agent-causal event (S's causing e), nothing causes (S's causing e) as a whole—causal relations themselves never get caused like that. Instead, the only possibilities

124. O'Connor, *Why Agent Causation?*, *supra* note 92, at 147.

> are that an event E or an agent A causes agent S, the first component of the complex whole (S's causing e). But these *prima facie* possibilities are not real, because S the agent is a substance and therefore cannot be caused by anything. S's coming to exist (an event) can be caused; S's changing (an event) can be caused; and S's ceasing to exist (an event) can be caused; but S (the substance itself) cannot be caused. Conceptually, agents are not the sorts of things that can be caused any more than touchdowns are the sorts of things that can be scored in chess matches. So everything requiring a causal explanation is accounted for—in particular, there are no uncaused events leftover. Agent causalism provides as much of an explanation of (S's causing e) as there is to have.

The critic might insist at this point that although event-causal events are never caused as a whole, agent-causal events are, indeed, must be, so caused. And because agent-causal events, as wholes, require causal explanations, the infinite regress is generated after all.[125] But this appears to be an *ad hoc* claim about causation. What motivates such a claim? Why think that causation is radically *discontinuous* in that way? For example, if the critic rejects agent causalism because of agent causalism's, allegedly, indefensible commitment to two fundamentally different sorts of causation, then why would the critic be satisfied emphasizing that there is a fundamental difference between sorts of causally complex events—between event-causal events and agent-causal events? What non-question-begging argument is available to the critic to defend such a fundamental difference? No such argument seems to be in the offing.

125. *Id.* ("[T]he claim is that events can directly bring about causal relations when they relate an agent to an event, but not when they relate an event to a further event."). Perhaps another implication of the critic's main claim here is that agents can directly bring about causal relations when they relate themselves to an event but not when they relate an event to a further event. The response to the critic would proceed the same way, *mutatis mutandis*.

The second of the more engaging objections to agent causalism that I shall consider tries to discredit agent causalism's ability to account for how agents can act for reasons and how reasons can play an indispensable role in explaining what agents do. In such reason-explanation accounts, "[a]n agent acts for a certain reason . . . only if the agent's recognizing that reason causes, in an appropriate way, the agent's behavior; and citing a reason contributes to a (true) reason-explanation of an action only if the agent's recognizing that reason caused, in an appropriate way, the action."[126] But the agent causalist cannot think that an agent's recognizing a reason alone is what causes the agent's free behavior, for the agent's recognizing a reason is an event, not a substance.[127] Instead, the agent causalist insists that the agent herself does at least some of the causing. So the agent causalist cannot provide a reason-explanation account in the usual way. Agent causalism is therefore faced with a challenge: Since the power of an agent to agent-cause events is reasons-responsive, agent causalism must be able to account for how reason-explanations of an agent's behavior are possible. But the standard way to do this is foreclosed to the agent causalist. So how can agent causalism supply the requisite account?

To respond to this challenge, the agent causalist must provide a plausible, non-standard account of reason-explanations. In Part V, drawing heavily on the work of others, I shall sketch such an account. My goal will not be to establish dispositively that the account is true but to establish that the account is plausible—that it might well be true. If the account to come is at least plausible, then agent causalism survives the second objection in that agent causalism remains plausible. I turn now to the task of providing that account by examining the anatomy of a voluntary act and situating a voluntary act within the context of the commission of a crime.

126. Randolph Clarke & Justin Capes, *Incompatibilist (Nondeterministic) Theories of Free Will*, *in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY (rev ed. 2013), http://plato.stanford.edu/entries/incompatibilism-theories/.

127. *Id.* ("Standard agent-causal views deny that events such as the agent's recognizing certain reasons cause any free action (or whatever event the agent directly causes when she acts freely).").

V. The Anatomy of a Voluntary Act and a Voluntary Act as a Part of a Crime

To return to the VAR: "[a] person is not guilty of an offense unless his liability is based on conduct which includes a voluntary act or the omission to perform an act of which he is physically capable."[128] In general, "if . . . the court can find a voluntary act by the defendant, accompanied at that time by whatever culpable *mens rea* that is required, which act in fact proximately causes some legally prohibited state of affairs, then the defendant is *prima facie* liable for that legal harm."[129]

The VAR is an aspect of the broader *actus reus* requirement, according to which there can be no criminal liability in the absence of an *actus reus*—"[t]he wrongful deed that comprises the physical components of a crime and that generally must be coupled with *mens rea* to establish criminal liability."[130] To connect the two legal requirements, "[t]he 'voluntary act' required for criminal liability is . . . to be understood as a bodily movement, and an *actus reus* can be analyzed into a (set of) bodily movements, and certain specified circumstances and consequences."[131] Thus, for the *actus reus* requirement to be satisfied, there must be, at the core of the *actus reus*, one or more voluntary body movements, which bring about prohibited states of affairs under specified circumstances. Oliver Wendell Holmes described this core as "a voluntary muscular contraction, and nothing else. The chain of physical sequences which it sets in

128. MODEL PENAL CODE § 2.01(1) (1962). Along similar lines in the context of a drunk driving charge:

> Though movement of a vehicle is an essential element of the statutory requirement, the mere movement of a vehicle does not necessarily, in every circumstance, constitute a 'driving' of the vehicle . . . . If a vehicle is moved by some power beyond the control of the driver, or by accident, it is not such an affirmative or positive action on the part of the driver as will constitute a driving of a vehicle within the meaning of the statute.

State v. Taft, 102 S.E.2d 152, 154 (W. Va. 1958).

129. *See* MOORE, ACT AND CRIME, *supra* note 16, at 35–36.

130. *Actus Reus*, BLACK'S LAW DICTIONARY 37 (7th ed. 1999).

131. Duff, *Acting, Trying, and Criminal Liability*, *supra* note 8, at 81–82.

motion or directs to the [resulting] harm is no part of it, and very generally a long train of such sequences intervenes."[132] The chain of consequences leads to or constitutes the state of affairs that is prohibited.

For example, "[a] person is guilty of criminal homicide if he purposely, knowingly, recklessly or negligently causes the death of another human being."[133] Imagine that A purposely kills B by shooting B. B's being shot to death is the prohibited state of affairs; A's moving her trigger finger is the voluntary act that causes the prohibited state of affairs; and the *mens rea* requirement[134] is satisfied because A purposely kills B—it was A's conscious objective to cause B's death. A has committed criminal homicide. As this example illustrates, the voluntary act is A's moving a part of her body, which kicks off a subsequent causal chain leading to the prohibited outcome. John Austin elaborates:

> Most of the names which seem to be names of acts, are names of acts, coupled with certain of their consequences. For example, [i]f I kill you with a gun or pistol, I shoot you: And the long train of incidents which are denoted by that brief expression, are considered . . . as if they constituted an act, perpetrated by me. In truth, the only parts of the train which are my act or acts, are the muscular motions by which I raise the weapon; point it at your head or body, and pull the trigger. These I will. The contact of the flint and steel; the ignition of the powder, the flight of the ball towards your body, the wound and subsequent death, with the numberless incidents

132. OLIVER WENDELL HOLMES JR., THE COMMON LAW 83-84 (Belknap Press 2009) (1881) (discussing acts in a torts context).

133. MODEL PENAL CODE § 210.1(1) (1962).

134. Note that the *mens rea* requirement corresponds to culpability, one of the factors that can influence a criminal defendant's desert according to standard retributivist theories. The MPC defines the four main types of *mens rea*—"purposely" (corresponding to purpose); "knowingly" (corresponding to knowledge); "recklessly" (corresponding to recklessness); and "negligently" (corresponding to criminal negligence). MODEL PENAL CODE § 2.02 (titled "General Requirements of Culpability.").

> included in these, are consequences of the act which I will. I will not those consequences, although I may intend them.[135]

The only things that get willed in the passage above are muscular body movements. Everything else is at most intended. And the "names of acts" to which Austin refers correspond to descriptions of what the agent does. Joel Feinberg suggests a metaphor for these act-descriptions:

> This well-known feature of our language, whereby a man's action can be described almost as narrowly or as broadly as we please, might fittingly be called the "accordion effect," because an action, like the folding musical instrument, can be squeezed down to a minimum or else stretched way out. He turned the key, he opened the door, he startled Paul, he killed Paul—all of these things we might say that Peter *did* with one identical set of bodily movements.[136]

Thus, two things are true of Austin's *pistol-shooting* finger movement. First, it is a muscular movement that initiates a causal chain; and second, it is the object, or part of the object, of a set of descriptions exhibiting the accordion effect. In addition, there are two other facts about the finger movement:

> A. The movement is, or is part of,[137] a basic action.
>
> B. The movement is voluntary.

135. JOHN AUSTIN, LECTURES ON JURISPRUDENCE OF THE PHILOSOPHY OF POSITIVE LAW 427–28 (Robert Campbell ed., 4th ed., London: John Murrary, Albemarle Street 1873) (emphasis removed).

136. Joel Feinberg, *Action and Responsibility*, *in* DOING AND DESERVING: ESSAYS IN THE THEORY OF RESPONSIBILITY 119, 134 (Princeton Univ. Press 1970) [hereinafter Feinberg, *Action and Responsibility*].

137. *See infra* for an explanation that the bodily movement will be one component of a basic action. It will not be identical to the basic action as a whole.

A. *The Movement is a Basic Action*

Actions can be divided into two kinds—basic and complex. A complex action is performed by performing some other action. For example, A prepares dinner by doing a number of other things—slicing vegetables, turning on the oven, and so forth. So preparing dinner is a complex action. In contrast, "*B* is a *basic action* of *a* if and only if (i) *B* is an action and (ii) whenever *a* performs *B*, there is no other action *A* performed by *a* such that *B* is caused by *A*."[138] For example, me wiggling my toe is a basic action because when I wiggle my toe there is nothing else that I do that causes my toe to wiggle.

Douglas Lavin elaborates the notion of a basic action:

> *Basic action* is a limit on [a] rational order of means and ends. It can be described from either side of the relation. Through the concept of an *end*: a basic action is not the end of any other action; nothing else is done in order to do it; it is not an answer to "Why?" when asked about any other action. And equally through the concept of a *means*: no means are taken in the execution of a basic action; it is not done by doing anything else; there is no answer to "How?" when asked of it. I illuminated the room by means of turning on the light, turned on the light by flipping the switch, and flipped the switch by moving my finger, but maybe moving my finger is something I simply did, something which did not involve taking any steps or means, or again doing anything with a view to moving my finger? If so, it is a basic action: *That X is doing/did A is basic just when there is no A* such that X is doing/did A* in order to do A; or again, That X is doing/did A is basic*

138. Arthur Danto & Sidney Morgenbesser, *What We Can Do*, 60 J. PHIL. 435, 435 (1963). I would suggest a third condition that must also be satisfied: (iii) whenever *a* performs *B*, there is no other action *A* performed by *a* such that *a* knows that *B*'s occurring is causally necessary for (and precedes) *A* and *a* performs *A* to assure *B*'s occurring.

> *just when there is no A* such that X is doing/did A by means of doing A*.*[139]

If A raises her arm, then A performs a basic action because there is nothing else that A does that causes her arm to rise. Of course, in the normal type of arm-raising case events happen before A's arm rises that cause A's arm to rise—bioelectrical impulses travel along A's arm-nerves and so forth—but these events are not actions[140] performed by A. Also, it is possible for A's raising her arm not to be a basic action. Imagine that A's arm is paralyzed but connected to a pulley system that, if activated, will raise it. In such a case, A's raising her arm by activating the pulley system would be a complex action, not a basic one.

B. *The Movement is Voluntary*

In addition to being a basic action, the *pistol-shooting* finger movement must be voluntary for the VAR to be satisfied. Beyond the idea, discussed earlier, that voluntariness guarantees a sort or degree of actor-control, what makes a body

139. Douglas Lavin, *Must There Be Basic Action?*, 47 NOÛS 273, 275 (2013) (emphasis added). Further:

> That there are such things as simple [i.e., basic] acts should be beyond controversy, partly because each person has direct experience of them in his own case and partly because a denial of their existence leads to an infinite regress and attendant conceptual chaos. If, before we could do anything, we had to do something else first as a means, then clearly we could never get started. As one writer puts it, 'If there are any actions at all, there are basic actions.'

Feinberg, *supra* note 136, at 136 (quoting Arthur Danto, *Basic Actions*, 2 AM. PHIL. Q. 141, 142 (1965)). *But cf.* Douglas Lavin, *Must There Be Basic Action?*, 47 NOÛS 273 (2013) (arguing that there might not be such things as basic actions).

140. One could conceptualize A's making her "arm-raising" arm-nerves fire as an action in certain situations. For example, imagine that A knows that a particular cluster of her arm-nerves fires and causes her arm muscles to contract every time she raises her arm, and assume that A sets out to fire those nerves by raising her arm. In such a case, A's firing her arm-nerves would be a complex action because condition (iii) (for basic actions) would not be satisfied. *See supra* note 138 and accompanying text.

movement a *voluntary* act in the sense of the VAR? The MPC provides the beginning of an answer, since it clarifies that body movements do not count as voluntary acts when they are "not a product of the effort or determination of the actor, either conscious or habitual."[141] The MPC also explicitly rules out certain categories of body movements from being voluntary actions, including reflex movements, unconscious movements, and movements resulting from hypnotic suggestion.[142]

Reconsider the MPC's positive characterization of voluntary actions, such as simple body movements, as those resulting from the "effort or determination" of the actor.[143] The MPC comments explain that bodily movements that result from an actor's effort or determination are to be understood as movements that result from an unimpeded exercise of the actor's will:

> [The MPC] formulates a residual category of involuntary movements, describing them as those that "otherwise . . . [are] not a product of the effort or determination of the actor, either conscious or habitual." The formulation seeks to express the main content of the traditional idea of an "external manifestation of the actor's will" . . . . In other respects the formulation . . . is designed to make the requirement of an act a minimal one.[144]

To flesh out the idea that a bodily movement is an "external manifestation of the actor's will," I shall adopt the view that a

141. MODEL PENAL CODE § 2.01(2)(d) (AM. LAW. INST., Official Draft and Revised Comments 1985).

142. *See* MODEL PENAL CODE § 2.01(2)(a)–(c) (AM. LAW. INST., Official Draft and Revised Comments 1985).

143. *See* MODEL PENAL CODE § 2.01(2)(d) (AM. LAW. INST., Official Draft and Revised Comments 1985).

144. MODEL PENAL CODE § 2.01 cmt. 2 (AM. LAW INST., Official Draft and Revised Comments 1985). Similarly, the VAR, "stated in its simplest form, is that the 'act' of the accused, in the sense of a muscular movement, must be willed. It must be a voluntary expression of the accused's will. . . ." H.L.A. HART, *Acts of Will and Responsibility*, *in* PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 90, 94–95 (2d ed. 2008) (quoting J. Ll. J. Edwards, *Automatism and Criminal Responsibility*, 21 MOD. L. REV. 375, 380 (1958)).

voluntary basic action is, at least in part, the causal result of the actor's preceding mental states—intentional mental states, in particular. (Here, "intentional" means having the property of intentionality (roughly, "aboutness").) The content of an intentional mental state is often understood to be a proposition. For example, Fred's belief that his room is clean has intentionality, and its content is the proposition that Fred's room is clean. If Fred has a phobia of clean rooms, then he might fear that his room is clean. If so, then his fear has intentionality, and its content is the same proposition—that Fred's room is clean. If Fred resolves to clean his room, forming the intention, or volition,[145] to do so, then his intention has intentionality, and again, its content is the same proposition—that Fred's room is clean. Because they have propositions as their contents, mental states such as beliefs, fears, and volitions are often referred to as *propositional attitudes*.

When someone acts voluntarily, what different sorts of intentional mental states or faculties are involved? The following is a partial breakdown of voluntarily visiting a neighbor:

> [I]f someone desires to visit his neighbour and believes that knocking on the neighbour's door will facilitate a visit, and hence forms the intention to knock on the neighbour's door, he is thereby prepared to exercise his will—he then tries to knock on the door, and with luck succeeds in performing the intended action.[146]

Featured in this breakdown are: (1) belief, (2) desire, (3) intention, and (4) will (the exercise of which is a "trying"). Turning first to (1)–(3): An intention is more like a desire than a belief in that desires and intentions are examples of what might be termed *pro-attitudes*, whereas a belief is not a pro-attitude. What does it take for a propositional attitude to be a pro-attitude? Not much. Pro-attitudes include "desires, wantings,

145. Although there is some dispute over the matter, I shall assume that a volition is an intention, not a type of belief or desire.

146. COLIN MCGINN, THE CHARACTER OF MIND: AN INTRODUCTION TO THE PHILOSOPHY OF MIND 131 (2d ed., Oxford Univ. Press 1982).

urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values in so far as these can be interpreted as attitudes of an agent directed toward actions of a certain kind."[147] Indeed, pro-attitudes might include:

> [N]ot only permanent character traits that show themselves in a lifetime of behavior, like love of children or a taste for loud company, but also the most passing fancy that prompts a unique action, like a sudden desire to touch a woman's elbow. In general, pro attitudes must not be taken for convictions, however temporary, that every action of a certain kind ought to be performed, is worth performing, or is, all things considered, desirable. On the contrary, a man may all his life have a yen, say, to drink a can of paint, without ever, even at the moment he yields, believing it would be worth doing.[148]

In a nutshell, a pro-attitude is any propositional attitude that favors the *coming true* of the proposition, that is, the attitude's content, and thereby can be connected causally to some action of the agent who has the attitude. Beliefs are not pro-attitudes even in this very broad sense, as they characteristically represent a proposition *neutrally*, as already being true.[149]

Pro-attitudes also have a characteristic *direction of fit*, which differs from the *direction of fit of a belief*. "It is characteristic of [beliefs] to represent the world as being a certain way, and [a belief] can be judged correct or incorrect according to whether the world is the way it is represented to be; the role of [beliefs] is to *fit* the world."[150] A mis-fitting belief is

147. Donald Davidson, *Actions, Reasons and Causes*, 60 J. PHIL. 685, 686 (1963).

148. *Id.*

149. I am simplifying by glossing over issues that might be raised, for example, by versions of moral internalism according to which believing certain propositions is itself sufficient for some degree of motivation.

150. MCGINN, *supra* note 146, at 117.

defective, and we call that defect *being false*. On the other hand:

> [D]esires [and other pro-attitudes] are said to have a "direction of fit" . . . that is the opposite to the "direction of fit" of beliefs . . . . [B]eliefs are like declarative sentences, which are satisfied (made true) by whether the world as it is conforms to them. But desires are like imperative sentences, which are satisfied (fulfilled) by changes in the world bringing the world into conformity with them.[151]

Although both intentions and desires are pro-attitudes, they are pro-attitudes of different types:

> [S]uppose I desire a milk shake for lunch, recognize that the occasion is here, and am guilty of no irrationality. Still, I might not drink a milk shake; for my desire for a milk shake still needs to be weighed against conflicting desires—say, my desire to lose weight. My desire for a milk shake potentially influences what I do at lunchtime. But in the normal course of events I still might not even try to drink a milk shake.
>
> In contrast, suppose that this morning I formed the intention to have a milk shake at lunch, lunchtime arrives, my intention remains, and nothing unexpected happens. In such a case I do not normally need yet again to tote up the pros and cons concerning milk-shake drinking. Rather, in the normal course of events I will simply proceed to execute (or anyway, try to execute) my intention and order a milk shake. My intention will not merely influence my conduct, it will control it.[152]

151. Tim Schroeder, *Desire, in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY (rev. ed. 2015), http://plato.stanford.edu/entries/desire/.

152. MICHAEL E. BRATMAN, INTENTION, PLANS, AND PRACTICAL REASON 15–

Thus, intentions are committed to action in a controlling, executory way that desires are not. Desires can influence, but "[a]s a conduct-controlling pro-attitude my intention involves a special commitment to action that ordinary desires do not."[153]

Having considered belief, desire, and intention in more detail, we can now ask: How do the actor's will and *trying* fit into a case where someone voluntarily performs a basic action such as raising her arm? Colin McGinn offers the following suggestion:

> [I]ntention cannot be analysed in terms of desire and/or belief, and willing cannot be reduced to intending . . . . Desire is unfettered by knowledge of what is practically possible, but intention needs to reckon with the practical facts of life, as these are seen by the agent. Intending is what channels desire and belief toward the will; forming an intention is like putting the active faculty into gear, without yet depressing the accelerator. But intending is not the *same* as willing. . . . [Y]ou can intend to do what you do not, in the event, will to do: you may intend to put a question to the distinguished speaker, but lose your nerve (will) at the last minute, though the intention may survive. We can say . . . that to will something is for the state of intending to be 'activated' . . . for an intention to be activated is just for the agent to try to do what he intends . . . . Without the will, then, intentions would never get off the ground. So the transition from . . . (desire and belief) to intention and thence to trying is a transition to genuinely distinct mental states or events, progressively closer, temporally and conceptually, to bodily action.[154]

16 (Harvard Univ. Press 1999).

153. *Id.* at 16.

154. MCGINN, *supra* note 146, at 131-32. McGinn's picture seems, in principle, empirically falsifiable. If we learn enough about how the human brain works (through scientific empirical investigation) such that it becomes

Two points about the trying aspect of a basic voluntary action deserve emphasis: (a) often, the trying is conscious to the agent, and (b) the trying is part of, instead of being the cause of, the basic action. Turning first to a trying being conscious, the notion of an unconscious trying seems natural when the action being attempted is a complex action:

> [T]here are descriptions of what an agent is trying to do that the agent is unaware of—as when a psychoanalyst says that his patient, in losing a photograph of his father, was trying unconsciously to get rid of his father. In such a case, we are describing the agent's trying in terms of his reasons,[155] and *these* can be unconscious.[156]

But often, when the action is a basic action, the agent is conscious of trying to perform it. When A tries, in the normal way, to raise her arm, A is often aware that she is trying to do so. Of course, there are cases in which A tries to raise her arm without being aware that she is trying, especially when A successfully and almost effortlessly tries. For example, imagine that someone asks A to raise her arm. A then tries to raise her arm and succeeds. A is aware of the request and of complying with the request. But A might not, in addition, be distinctly and consciously aware of trying to comply with the request. To A, it may seem as if she just effortlessly does what she was asked to do. But even if in such a case A is not distinctly conscious of trying, A is nonetheless trying—there is a trying going on in addition to the arm movement. In other words, every basic action includes a trying, even when the action is successful and thereby draws the actor's attention away from the *trying* part of what she does.

implausible to think that a human brain could enter into functional states that satisfy McGinn's picture of how mental states cause voluntary actions, then the accuracy of the picture would be called into serious question.

155. I would be inclined at this point in the passage to refer to unconscious *desires*, instead of *reasons*, to explain (without evaluating) the agent's behavior.

156. MCGINN, *supra* note 146, at 128.

As previously mentioned, the trying is also part of the basic action—the trying does not cause the action. When A raises her arm: (i) A tries to raise her arm, and (ii) it rises. Both the trying and the movement are components of the action:

> [T]he trying occurs, closely followed by the movement, these being related . . . as cause and effect; the [basic] action is . . . trying and movement *taken together*. More precisely, the action is . . . composed of both the trying and the movement—or equivalently, is . . . identical with a complex event having these constituents. Since the action has these two items as constituent components, it is neither caused by the trying nor the cause of the movement—for causal relations do not hold between events and their constituents.[157]

In other words: The trying causes the arm to rise. But the trying does not cause the basic action, and the basic action does not cause the arm to rise. Instead, the trying and the arm's rising compose the basic action. All voluntary basic actions include, without being caused by, a trying. At this point, most of the main elements of the anatomy of a voluntary basic action have been identified—belief, desire, volition, and trying (willing). What remains is to insert the agent and then explain how the account might accommodate the reasons-responsiveness of agent-causation.[158] But before doing this in Part V.D, it is worth considering some of the advantages of including a trying as a part of a voluntary basic action. These advantages indirectly bolster the claim that there are plausible versions of agent causalism.

C. *Advantages to Understanding Voluntary Basic Actions to Include Tryings*

157. *Id.* at 126.

158. *See supra* Part IV. Recall that accommodating reasons-responsiveness was the challenge posed to agent causalism at the end of Part IV.

Adopting the position that a voluntary basic action always includes, as a distinct element, an exercise of will (a trying) has advantages. R.A. Duff poses a hypothetical and broaches a question famously raised by Ludwig Wittgenstein.[159] First, the hypothetical:

> Whether I move my arm depends on conditions outside my control: if those conditions are not satisfied (if my arm is paralyzed . . .), I might fail to move my arm, although . . . I try to move it. I might even believe that I have moved my arm when it has not actually moved: this happens when someone's arm has been anaesthetized, he is asked to shut his eyes and raise his arm, and the arm is held down. This person has not been merely inactive.[160]

Wittgenstein's famous question was: "[W]hen 'I raise my arm,' my arm goes up. And the problem arises: what is left over if I subtract the fact that my arm goes up from that fact that . . . I raise my arm?"[161]

How do we explain the active nature of the person with the anaesthetized arm? And is there a way to answer Wittgenstein's question that is related to how we explain that person's active nature? One straightforward way to do this is to say that the person was trying to raise his arm (trying to perform a basic action) and that what is *left over* when I subtract the fact that my arm goes up from the fact that I raise my arm is my trying to raise my arm.

159. Duff ultimately argues against the idea that every basic action includes a distinctive trying. *See* Duff, *Acting, Trying, and Criminal Liability, supra* note 8, at 75.

160. *Id.* at 83-84. One could modify the hypothetical so that the arm is not being held down and that the individual, without realizing it, has been administered a temporary *arm paralytic*. Or maybe the anesthetic doubles as a paralytic, but the individual does not realize that. Even in such a modified case, the individual is not *merely inactive* in the sense under consideration here.

161. LUDWIG WITTGENSTEIN, PHILOSOPHICAL INVESTIGATIONS para. 621 (G. E. M. Anscombe trans., Basil Blackwell 1963).

In Duff's example of the person who tries to raise his arm but fails, without realizing it, there is no basic action because there is no arm movement. In the normal arm-raising case, there is the trying and the arm movement. The trying causes the arm movement. And the voluntary basic action comprises the trying causing the arm movement. In the abnormal case of the man with the anesthetized arm, there is no basic action because only the trying happens—the arm movement part is missing because something interferes with the causal link between the trying and the arm.

Consistent with Part V.B's brief discussion of trying, and limiting our attention for the moment to basic actions, the following picture emerges:

> [A]ll [basic] actions . . . involve trying . . . . Trying is inherently active, and . . . [is] the psychological aspect of action. We do not, of course, always *say* of someone who acts that he tried to do that which he did . . . . But this does not imply that it is *false* to claim that agents try to perform even their most effortless actions . . . . [T]here seems no difference, with respect to what is going on in you psychologically, between the normal case in which your arm rises as a result of your decision to raise it and the abnormal case in which, unknown to you, your arm has been paralyzed; yet in the latter case we would say that you did at least try to raise your arm—and your mental acts were no different in the former case . . . . [S]o we can legitimately claim that there is always an event of trying involved in any [even basic] action.[162]

Adopting the view that tryings are distinct elements of every voluntary basic action also helps explain certain types of *voluntary* omissions. Sometimes omissions have an agent-controlled, voluntary nature. "[A] guardsman who keeps himself from moving acts, but acts precisely by not moving his body."[163]

162. MCGINN, *supra* note 146, at 123-24.

163. Duff, *Acting, Trying and Criminal Liability*, *supra* note 8, at 83.

The active, voluntary nature of the guardsman is captured by the idea that he is, successfully, trying to keep his body still—trying to not-move his body. Tryings serve not just as the psychologically active components of voluntary actions but also as the psychologically active components of voluntary omissions.[164] As explained, in the case of a voluntary arm-raising, the trying and the arm's rising are two components of the basic action, and the trying causes the arm movement. In the case of a voluntary non-raising of an arm, the trying and the arm's not rising are two components of the voluntary omission, and the trying causes the arm's not rising. (Contrast this last case with the previous case of the anesthetized arm that does not rise. In the anesthetized arm case where the actor is trying to raise his arm, there is the trying and the arm's not-rising. But there, the trying does not cause the arm's not-rising. Instead, the causal explanation of the arm's not-rising is that the anesthesia/paralytic interferes with the usual causal connection between what the actor tries to do with his arm and how his arm behaves. So in the anesthetized arm case, there is an omission (to raise the arm), but the omission is not voluntary.)

Adding tryings to the picture—distinct from beliefs, desires, and volitions—also helps with what might be referred to as the problem of wayward causal chains. Donald Davidson raised this problem against the backdrop of a theory according to which it is an agent's belief/desire pair that proximately causes body movements when she intentionally acts. Davidson noted that, even if a belief/desire pair represents and causes a simple muscular movement, there is no guarantee that what happens is a voluntary, intentional action:

> Beliefs and desires that would rationalize an action if they caused it in the *right* way . . . may cause it in other ways. If so, the action was not performed with the intention that we could have read off from the attitudes that caused it . . . . A

164. This is not to suggest that an omission must be voluntary (include a trying) for the VAR to be satisfied. For example, if the requisite *mens rea* is (criminal) negligence and the *actus reus* is defined in terms of an omission, then the VAR might be satisfied even if the defendant does not actively try to omit.

> climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally.[165]

Arguably, the VAR might not be satisfied if criminal charges were brought against the climber for the death of his climbing partner.[166] Granted, his grip-loosening was not an entirely mindless reflex, an unconscious movement, or the result of hypnotic suggestion. But something was not voluntary about what the climber did. An official comment to the MPC emphasizes that "[t]here is sufficient difference between ordinary human activity and a reflex or a convulsion to make it desirable that they be distinguished for purposes of criminal responsibility by a term like 'voluntary.'"[167] Voluntary actions are supposed to be instances of "ordinary human activity." To assure that we have a genuine case of "ordinary human activity," we should stipulate that the belief/desire pair causes the muscular movement in the "right way"[168] and not via a wayward causal chain.

But what is the "right way?" Once we introduce tryings, it becomes natural to say that the VAR is not satisfied in the mountain-climber case because the climber never tried to let go. Causal chains are prevented from becoming wayward by having tryings situated within them. Of course, this raises another

165. Donald Davidson, *Freedom to Act*, *in* ESSAYS ON ACTIONS AND EVENTS 63, 79 (2d ed., Oxford Univ. Press 1980).

166. I do not mean to suggest that criminal charges would be brought in a case like this.

167. MODEL PENAL CODE § 2.01 cmt. 1 (AM. LAW INST., Official Draft and Revised Comments 1985).

168. *See* Bishop, *supra* note 108, at 61. Obviously, a defensible specification of what counts as the "right way" is necessary to solve the "wayward causal chain" problem and clarify exactly what makes for a voluntary act. See also *id.* at 61-79 for an argument that the most promising strategy for solving the problem of wayward causal chains is to adopt agent causalism.

question. How are tryings situated within causal chains to solve the problem of wayward causal chains and assure that the VAR is satisfied? To answer this question, I turn now to completing the sketch of the anatomy of a voluntary action. Doing this will suggest how tryings might solve the problem of wayward causal chains, and it will also suggest an answer to the remaining open question from Part IV: How might agent-causation be reasons-responsive?

D. *How Agent-Causation Might be Reasons-Responsive*

How can we fit agents into our developing picture of voluntary actions to make agent-causation responsive to reasons? To answer this question, perhaps a good way to begin is to reflect on what it is like to make an everyday decision voluntarily to act on the basis of conscious deliberation—on the basis of practical reasoning:

> When I decide . . . to go for a walk on a cool autumn evening, I am conscious of various factors at work . . . motivating me either to do so or to do something else instead. And there are some courses of action which, while it is *conceivable* that I might choose to follow them . . . do not represent 'genuine' possibilities for me at that time, given my current mood, particular desires and beliefs, and, in some cases, long-standing intentions of a general sort. But within the framework of possibilities . . . that these . . . conative and cognitive factors set, it seems . . . to be *up to me* to decide which particular action I will undertake. The decision I make is no mere vector sum of internal and external forces acting upon me during the process of deliberation . . . . Rather, *I* bring it about—directly, you might say—in response to the various considerations: I am the source of my own activity, not merely in a relative sense as the most proximate and salient locus of an unbroken chain of causal transactions leading up to this event, but fundamentally, in a way not

prefigured by what has gone before.[169]

This passage partially describes what it is like to decide to voluntarily go for a walk as the result of deliberation—as the result of practical reasoning. And it mentions a causal role that the agent herself plays in making that decision. Further, the ultimate output of the agent's deliberation is a physical action that the agent performs—going for a walk. An ancient question is impressed upon us when we consider that physical actions are the ultimate outputs of episodes of practical reasoning:

> [H]ow is it that thought . . . is sometimes followed by action, sometimes not; sometimes by movement, sometimes not? What happens seems parallel to the case of thinking and inferring about the immovable objects of science. There the end is the truth seen (for, when one conceives the two premisses, one at once conceives and comprehends the conclusion), but here the two premisses result in a conclusion which is an action—for example, one conceives that every man ought to walk, one is a man oneself: straightway one walks; or that, in this case, no man should walk, one is a man: straightway one remains at rest.[170]

This ancient question about practical reasoning arises largely because, in paradigmatic cases, practical reasoning's inputs are psychological and logical, but its outputs are strikingly different—they are muscular actions.

Practical reasoning is often distinguished from theoretical reasoning. Theoretical reasoning might be considered less mysterious than practical reasoning because its outputs do not differ as dramatically from its inputs. Both the inputs and the outputs are beliefs in cases of theoretical reasoning, which proceeds in the form of theoretical inferences. It is important to

169. O'Connor, *Agent Causation, supra* note 113, at 173.

170. ARISTOTLE, ON THE MOTION OF ANIMALS 1 (A. S. L. Farquharson trans., Infomotions, Inc. 2001).

distinguish: (a) drawing a theoretical inference and (b) a theoretical argument. Making an inference is psychological as well as logical. Drawing an inference is something that someone does; it is a psychological action. A theoretical argument, in contrast, is a set of propositions, some of which are premises and one of which is the conclusion. The relations between the propositions are not psychological; they are logical, or evidential. An argument is not something that someone does. Of course, making an argument is something that someone does; it is a communicative act. Also, when an argument is proffered and understood, there can be psychological effects, for example, the audience becomes convinced that the argument's conclusion is true.

Drawing a theoretical inference begins with the reasoner's believing that an argument's premises, at least, considered individually, are true. The causal upshot is the reasoner's coming to believe that the argument's conclusion is true. In theoretical inferences, the connection between the input and output beliefs is forged by the reasoner's grasp of the logical or evidential relationship between the premises and conclusion. To take a simple example, consider the following deductive syllogism:

(1) All tortoises are mortal.

(2) Socrates is a tortoise.

Therefore, (3) Socrates is mortal.

Imagine that the reasoner believes that (1) and (2) are true. John Bishop has suggested the following:

> When I infer that Socrates is mortal from my beliefs that Socrates is a tortoise and all tortoises are mortal, *I* do something. It is clear that, when I make the inference, it is not the case that my believing that Socrates is a tortoise and my believing that all tortoises are mortal are jointly causally sufficient for my believing that Socrates is mortal. When I make this inference, I consider the content of these two beliefs of mine, recognize their mutual relevance, and reason it out that Socrates is mortal. This is not a very difficult task, yet it is conceivable that I should not carry

> it out. I might fail to grasp the validity of the form of argument I need to use, or, more likely, though I do believe that Socrates is a tortoise and that tortoises are mortal it happens that I never connect the two: the tortoisehood of Socrates and the mortality of tortoises never come together in the same train of thought. Now, in this case, though we refuse to allow that my holding the premiss beliefs itself suffices causally for the formation of the conclusion belief, we do not deny that the premiss beliefs play some causal role. They are intrinsic to the inference which brings about the new belief. It is the agent who makes the inference, and so that agent who causes his belief, but he does so only in virtue of holding the premiss beliefs to be true.[171]

This passage suggests that, even in the case of coming to believe that Socrates is mortal by drawing a theoretical inference, the agent causes something. The agent causes her own coming-to-believe the conclusion; the agent herself causes an event. In other words, it seems that agent-causation might play a role not only in voluntary action, but also as a part of theoretical reasoning. Of course, the agent's premise-beliefs also play a crucial causal, and logical, role. We have a picture in which the agent herself is inserted as a causal factor, along and in concert with her intentional mental states. And the agent is exercising her agent-causal power, to draw an inference, in a reasons-responsive manner. In combination, the fact that all tortoises are mortal and the fact that Socrates is a tortoise are reasons to believe that Socrates is mortal. And when the agent draws the theoretical inference, she is responding to those reasons.

Even if agent-causation is operative in theoretical reasoning as well as in voluntary action, there is a certain degree of control that seems lacking in connection with reasoning, at least in comparison to one's control over one's basic muscular actions. Granted, I do seem to have control over the *trying* aspects of

171. Bishop, *supra* note 108, at 77-78.

performing mental actions, like drawing an inference. It is up to me how hard I try, on a given occasion, to recognize logical or conceptual relationships. But, I routinely seem to lack control over what I come to believe as a result of such efforts. (Contrast: I do not routinely lack control over where my arm goes as a result of trying to raise it.)

For example, imagine that I try to follow a proof of the Pythagorean Theorem and succeed. I see that the requisite logical and conceptual relationships hold between the proof's premises and conclusion, and I see that the premises are undoubtedly true. As a result, I come to believe the Pythagorean Theorem. But now I am, so to speak, stuck with the new belief whether I like it or not. Having drawn the inference, I now cannot help myself epistemologically. Unless my memory fades, I cannot rid myself of that new belief, no matter how badly I might wish that the Pythagorean Theorem were false. Empirical beliefs are also often *involuntary* in roughly this sense. I believe that there is a keyboard on which I am typing right now. Of course, I could contemplate bizarre falsifying hypotheses in an effort to call that belief into doubt. Maybe I am having a vivid dream and am in fact abed, far away from any keyboards. But this, at most, reduces my certitude a little. I just cannot get myself to not believe that I am typing on a keyboard now.

Bracketing this seeming difference between our control over our beliefs and our control over our voluntary basic muscular actions, Bishop's sketch of how reasons-responsive agent-causation is operative in theoretical reasoning suggests how such causation might fit into practical reasoning leading ultimately to voluntary action. Someone might understand a practical inference as having beliefs as inputs and an intention as an output. To take a simple example, imagine that A promises B that she will meet B at the bus station at 3:00 PM. Consider the following series of propositions:

(1) A promised B that she will meet B at the bus station at 3:00 PM.

(2) A should keep her promises.

(3) To get to the bus station by 3:00 PM, the only means available to A is to ride her bicycle.

(4) A rides her bicycle to the bus station.

How might A's drawing a practical inference work in this case? A believes (1), (2), and (3). That is, A has three beliefs, whose propositional contents are (1), (2), and (3), respectively. A recognizes that (1), (2), and (3), considered together, are practical reasons that count in favor of riding her bicycle to the bus station. A reasons it out and forms the intention to ride her bicycle to the bus station—A forms an intention whose propositional content is (4). This is not a very difficult task, yet it is conceivable that A not carry it out. A might fail to grasp the mutual relevance of the practical reasons. She might believe that she made the promise, believe that she should keep her promises, and believe that her only chance to get to the bus station by 3:00 PM is to ride her bicycle. But she might never connect the three in a unified train of thought. Although A's holding the premise-beliefs does not itself suffice causally for the formation of the intention to ride her bicycle to the bus station, the premise-beliefs play a causal role. They are intrinsic to the inference that brings about the intention. It is the agent who makes the inference, and so that agent who causes her intention, but she does so only in virtue of holding the premise-beliefs to be true.

Thus, reasons-responsive agent-causation can play an indispensable role in the formation of intentions (volitions). But this does not get us all the way to voluntary action, because "you can intend to do what you do not, in the event, will to do."[172] "Intending is what channels desire and belief toward the will; forming an intention is like putting the active faculty into gear, without yet depressing the accelerator."[173] To get all the way to voluntary action, the agent must do one more thing—depress the accelerator. In other words, the agent must try. And, as discussed previously, when the agent tries, she agent-causes the trying. Further, if the trying causes a bodily movement, then the agent also voluntarily agent-causes a basic action, composed of the trying and the bodily movement. And when the agent voluntarily performs the complex action of riding her bicycle to the bus station, she does so by voluntarily performing numerous

172. MCGINN, *supra* note 146, at 132.
173. *Id.*

basic actions.

The preceding paints a picture of voluntary action that the actor performs on the basis of practical reasoning. According to this picture, agent-causation is indispensable. First, agent-causation comes into play in the actor's reasons-responsive formation of a volition. Without the volition, there is no voluntary act. So the reasons-responsive agent-causation is essential to the provenance of the voluntary act. Second, because the volition, standing alone, is insufficient to cause the voluntary action, agent-causation kicks in a second time. The agent exercises her will—she tries. If the trying successfully causes the intended muscular movement, then the agent voluntarily acts—the voluntary basic act consists of the trying and the muscular movement. And all voluntary actions are either basic or they are done by performing basic muscular actions. This is how agent-causation ultimately leading to voluntary action might be reasons-responsive.[174]

This picture also suggests how tryings might solve the problem of wayward causal chains. The trying that solves the problem of wayward causal chains fits in after the formation of the intention. An agent's beliefs, desires, fears, etc. may influence what she does, even after she forms the intention. But the intention, being executory, puts the agent's will into gear, so to speak. And once that happens, it is up to the agent to exercise her will, (i.e., to agent-cause a trying). The intention does not deterministically event-cause the agent to exercise her will (i.e., the intention is not causally sufficient for her to try). When a trying is situated between the agent's mental states and her action in this way, the link between the mental states and the action is not wayward, and the action is voluntary.

VI. Agent-Causal Retributivism and Justifying Applications of the VAR

As this paper has argued, for retributivist justifications of the VAR to be plausible, agent causalism must be true. Parts IV

174. Thus, the challenge posed by the second objection to agent causalism at the end of Part IV has been met (at least, well enough that agent causalism emerges as plausible).

and V have been dedicated largely to making agent causalism plausible. But many think that agent causalism is false. And if it is false, then this paper's main thesis implies that retributivism cannot contribute to justifying the VAR—the fundamental predicate of legal criminal liability. Before concluding, it is worth considering whether retributivism would have difficulty justifying the VAR even if agent causalism were true. To do this, it is helpful to consider three kinds of cases recently discussed by Gideon Yaffe.

Yaffe has argued that the VAR is justified because it assures that what he terms the "Requirement of Correspondence" is satisfied when criminal liability is imposed upon a defendant:

> When a defendant is shown to be guilty of a crime, he is shown to have performed certain acts with certain results in certain circumstances . . . . And he is shown to have been in certain mental states . . . . But . . . there is an additional requirement that is so rarely at issue as to go unmentioned most of the time: the defendant's actions must correspond with his mental states.[175]

In brief, the Requirement of Correspondence is that the defendant's *mens rea* and *actus reus* must correspond in the right way for the state to impose legal punishment. And the VAR guarantees that this requirement is satisfied:

> Voluntary acts matter to criminal liability . . . because without them we lack the link between objectionable mental states and objectionable acts that is required to be justified in punishing for the package of mental states and conduct that crimes . . . consist in. There is *mens rea* and there is *actus reus*; but without a voluntary act, there is not the link between the two that is required for desert of punishment for the conjunction.[176]

175. Yaffe, *supra* note 2, at 183.
176. *Id.* at 184.

Yaffe refers to this way of substantiating the VAR as the "Manifestation of *Mens Rea* Rationale":

> Under [this rationale], the VAR is a byproduct of the idea that *mens rea* is an essential part of criminal liability. It is because we already think that people should not be punished in the absence of a showing of *mens rea* . . . that we are barred, for moral reasons, from punishing them in the absence of a voluntary act. *Mens rea* is essential, but it isn't relevant unless it's manifested. And it isn't manifested unless there's a voluntary act. To punish . . . in the absence of a voluntary act is morally no different from punishing in the absence of *mens rea*, and that is unacceptable.[177]

And Yaffe argues that the Manifestation of *Mens Rea* Rationale squares with three types of cases in which the VAR plays a decisive role.

A. *Cases Featuring Complex Unconscious Bodily Movements*

In the first type of case, there is no criminal liability because unconscious[178] bodily movements do not count as voluntary acts for the purposes of the VAR. As one court has clarified, "[t]o constitute a defense, unconsciousness need not rise to the level of coma or inability to walk or perform manual movements."[179] A striking example is provided by a leading Canadian case, *R. v. Parks*.[180] In *Parks*, the defendant drove his car about twenty-three kilometers from his residence to his in-laws' home.[181] He then attacked his in-laws while they were asleep, killing one of them.[182] Afterward, he drove his car to a nearby police station,

177. *Id.*

178. Some might distinguish "unconscious" mental states from "subconscious" mental states. This paper attempts no such distinction. Hereafter, I shall use only the term "unconscious."

179. People v. Halverson, 165 P.3d 512, 539 (Cal. 2007).

180. R. v. Parks, 2 S.C.R. 871 (Can. 1992).

181. *Id.* at 871.

182. *Id.*

telling the police what he had done.[183] The defendant was acquitted.[184] He successfully argued that because he was sleepwalking through the entire incident, he should not be subject to criminal liability.[185] On appeal, the Supreme Court of Canada determined that the defendant indeed should have been acquitted because the record indicated that he was in a state of "non-insane automatism" during the incident:

> Automatism, although spoken of as a "defen[s]e", is conceptually a sub-set of the voluntariness requirement, which in turn is part of the *actus reus* component of criminal liability. An involuntary act, including one committed in an automatistic condition entitles an accused to an unqualified acquittal, unless the automatistic condition stems from a disease of the mind that has rendered the accused insane.[186]

The court added:

> [Canada's] system of justice is predicated on the notion that only those who act voluntarily should be punished under the criminal law. Here, no compelling policy factors preclude a finding that the accused's condition was one of non-insane automatism.[187]

How might an agent-causal, retributivist justification of the VAR handle such cases? In the paradigmatic agent-causal situation outlined above: (i) the agent forms a volition; (ii) the volition enables the agent's will to become active; (iii) the agent

183. *Id.*

184. *Id.* at 872.

185. *Id.* at 871-82.

186. *Parks*, 2 S.C.R. at 872.

187. *Id.* at 874. *Cf.* Fain v. Commonwealth, 78 Ky. 183, 193 (1879) (granting the defendant a new trial because he had not been allowed to prove that he suffered from somnambulism) ("Our law only punishes for overt acts done by responsible moral agents. If the prisoner was unconscious when he killed the deceased, he cannot be punished for that act . . . .").

exercises her will—she tries; and (iv) assuming that her trying causes her body to move, she herself thereby causes an action. When (i)–(iv) happen, a defendant exercises the requisite control, which, in combination with the *mens rea* and the remainder of the *actus reus* being satisfied, makes her deserving of legal punishment.

The agent-causal retributivist justification of the VAR would account for cases where the defendant acts unconsciously if it turned out that in such cases the defendant lacked the control requisite for desert. Assuming that he was unconscious the entire time, did Parks lack that sort of control when he killed one of his in-laws? The agent causalist might argue that Parks lacked the requisite control because he did not, in the technical sense of the agent-causal picture outlined in Part V, *try* to do any of the things that he unconsciously did during his sleepwalking episode. This argument relies on the following principle:

> (P) The trying element of agent-causal voluntary action is missing in any case in which a defendant is unconscious of what she is doing.

Initially, this agent-causal explanation may seem very counterintuitive. (Surely, for example, Parks tried to drive to his in-laws' home. If he did not even try, how could he have managed to arrive there by car?) It is important to note, however, that (P) uses the term *trying* in a technical sense. As elaborated previously: a *trying* is a mental event distinct from other types of states such as beliefs, desires, fears, and volitions; a *trying* plays a particular functional role in the context of a normal voluntary action; a *trying* is an event that is agent-caused; a *trying* is a component of, but not the cause of, a basic action; and so forth. If we lose sight of this, then of course it may seem absurd to say that Parks did not even *try* to drive to his in-laws' home.

But with the technical meaning of *trying* firmly in mind, is (P) plausible? The issue here is not whether the trying itself is conscious. As previously explained, sometimes an actor is not distinctly conscious of trying to do X but instead is conscious only of doing X. (The previous example involved someone raising her

arm upon request.)[188] But as also explained, in such cases the actor is still trying to do X. (The actor was still trying to raise her arm.) But what if the actor is unconscious of doing X? (P) does not imply merely that the agent is also not conscious of trying to do X. According to (P), in such a case there is no trying for the agent to be conscious of.

If (P) is true—if there is no trying in cases, such as *Parks*, involving complex unconscious bodily movements—then how can the complexity of the bodily movements be accounted for? What Parks did was fairly elaborate. He got into his car; turned on the ignition; drove about twenty-three kilometers to a destination where two people he knew personally were located; and so on. That sort of behavior evinces rationality. The agent causalist could account for the complexity the same way an event causalist might. For example, Yaffe describes the conduct of Huey Newton, who defended himself in court by arguing that he was unconscious when he fatally shot a police officer:[189]

> Newton's finger movements on the trigger were not likely to be purely reflexive; they were clearly goal directed. Newton seems to have been *aiming* the gun at the officer, and so must have been mentally representing a particular goal, namely to shoot the officer, a mental representation that was involved in guiding his bodily movements.[190]

To explain what Newton did, we need to treat him as an intentional system. An intentional system is "a system whose behavior can be (at least sometimes) explained and predicted by relying on ascriptions to the system of beliefs and desires (and hopes, fears, intentions, hunches . . . ."[191] Intentional states, beliefs, desires, volitions, etc., play crucial roles in explaining how intentional systems behave. An agent causalist can avail herself of such states just as an event causalist can. An agent causalist could consistently hold that intentional states operate

188. *See supra* pp. 162-63.
189. *See* People v. Newton, 87 Cal. Rptr. 394 (Ct. App. 1970).
190. Yaffe, *supra* note 2, at 176.
191. Daniel Dennett, *Intentional Systems*, 68 J. PHIL. 87, 87 (1971).

deterministically in cases of complex unconscious behavior. In such cases, the agent herself is causally inert because she is *asleep*. But her rich, variegated, causally-complex economy of unconscious intentional states is not inert. Because the agent herself is inert, she lacks the control requisite for desert. In this way, an agent causalist retributive justification of the VAR could be consistent with not imposing criminal liability for complex unconscious conduct.

B. *Cases Featuring Certain Omissions*

In the second type of case, there is criminal liability even though there is no voluntary act because the defendant is guilty of a certain sort of omission. Certain omissions are equivalent to voluntary acts for the purposes of the VAR. A good example is provided by *People v. Manon*.[192] In *Manon*, the defendant's appeal from a conviction for criminally negligent homicide and endangering the welfare of a child was denied.[193] Her infant son had died due to neglect.[194] The court emphasized that "[c]riminal liability may . . . be based upon an omission, if such omission is the failure to perform a legally imposed duty such as parents' nondelegable affirmative duty to provide their children with adequate medical care."[195]

How might an agent-causal retributivist justification of the VAR handle such cases? When a crime's *actus reus* features an omission, there are two main possibilities depending on what sort of *mens rea* is required for the crime. If the requisite *mens rea* is purpose, then the agent causalist might insist that the *actus reus*'s omission, a state of affairs, be caused by the agent via a trying, as elaborated previously. Recall the case of the guardsman who actively refrains from moving.[196] In that case, the guardsman agent-caused his not-moving. If that sort of

192. People v. Manon, 640 N.Y.S.2d 318 (App. Div. 1996), *leave to appeal denied*, 673 N.E.2d 1248 (N.Y. 1996).

193. *Id.*

194. *See id.* at 319.

195. *Id.* at 320 (quoting People v. Steinberg, 595 N.E.2d 845, 847 (N.Y. 1992)).

196. *See supra* p. 166.

omission were a crime,[197] then perhaps the most natural *mens rea* would be purpose. The guardsman would be guilty of this crime because he purposely *not-moved*. In general, if a penal code defined crimes in terms of purposeful omissions, then the agent-causal retributivist would have no problem defending the VAR, which would be interpreted as requiring a trying before criminal liability was imposed. Only when the defendant tried to omit would she exercise the control necessary for moral responsibility, and therefore for desert, as least as to crimes of omission requiring that type of *mens rea*.

But if the *mens rea* of a crime of omission were knowledge, recklessness, or criminal negligence or if the crime of omission were a strict liability crime, then the defendant's trying would presumably no longer need to be the cause of the omission. Take, for example, the penal statutes that Cindy Manon violated when she neglected her infant son:

> A person is guilty of criminally negligent homicide when, with criminal negligence, he causes the death of another person. N.Y.[198]

> A person is guilty of endangering the welfare of a child when . . . . Being a parent . . . he or she fails . . . to exercise reasonable diligence in the control of such child to prevent him or her from becoming . . . a "neglected child" . . . . [199]

Section 125.10 required only criminal negligence for guilt.[200] And because Manon was a parent, § 260.10 did not require that she have any particular *mens rea*, beyond that possibly suggested by "reasonable diligence," to be guilty.[201] Manon was the victim's parent, and she had at least two legal duties toward her son—a duty to not negligently cause his death[202] and a duty

197. Of course, it is farfetched to think that such an omission would ever actually be a crime.

198. N.Y. PENAL LAW § 125.10 (McKinney 2016).

199. N.Y. PENAL LAW § 260.10.

200. *See* N.Y. PENAL LAW § 125.10.

201. *See* N.Y. PENAL LAW § 260.10.

202. Manon's having this particular duty did not depend on her being the

to exercise reasonable diligence to prevent him from being neglected. Manon flouted both duties. And her flouting them had nothing in particular to do with her exercise of an agent-causal ability to cause crucial omissions. So how might an agent-causal retributivist justify the law's position that the VAR is satisfied in such cases?

To see how, first consider Yaffe's explanation of how the Manifestation of *Mens Rea* Rationale addresses such omissions:

> [W]hat we *do not do* often manifests our objectionable mental states just as much as what we *do do* even if there is no volition present. The mental state of disregarding one's child's welfare is manifested by *the failure* to do that which the child's welfare requires that one do. However, in such cases, there need be no volition to serve as causal intermediary between the morally and criminally relevant mental state— . . . the state of "disregard"—and the failure to do as one ought. While that failure may need to be caused by the prior mental state for the failure to manifest itself in the morally relevant way, such causation does not require volitional intermediaries.[203]

The mental state featured in Yaffe's account is a "state of disregard."[204] (We can put knowledge and recklessness aside for the moment, since disregarding one's child does not require knowledge or recklessness.) Thus, Yaffe seems to treat negligence, or perhaps even strict liability, as a type of mental state—it seems that disregarding negligently, or disregarding, period, is supposed to count as being in a type of mental state. This seems reasonable, since negligence and strict liability can be considered culpability ("*mens rea*") states, and "mens rea" translates to "guilty mind."[205] But unlike recklessness, knowledge, or purpose, which require conscious psychological states, negligence and strict liability do not:

parent of the victim.

203. Yaffe, *supra* note 2, at 187–88.

204. *Id.*

205. *Mens Rea*, BLACK'S LAW DICTIONARY 999 (7th ed. 1999).

> Properly understood, the principal mental state concepts do not reflect a single hierarchy of legal significance. Rather, they conceal two distinct mental state hierarchies, of desire and belief, as well as a third category, of conduct, which does not essentially involve mental states. . . . Culpable conduct includes . . . gross negligence.[206]

In other words, negligence and strict liability are types of *mens rea*, but they are not types of mental, in the sense of *psychological*, states. For example, the objective, counterfactual nature of the MPC's definition of negligence reinforces that negligence is not a mental state:

> A person acts negligently with respect to a material element of an offense when he should be aware of a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that the actor's failure to perceive it, considering the nature and purpose of his conduct and the circumstances known to him, involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation.[207]

Mental, psychological, states are at most indirectly involved or implied—the *purpose* of the defendant's conduct and the defendant's *knowledge* of surrounding circumstances. But these serve only to elaborate aspects of a hypothetical situation that a *reasonable person* is placed in. Regardless of a defendant's actual psychological states, she is negligent when she fails to do what a reasonable person would do. And a standard definition of a strict liability crime reinforces that strict liability is not a

206. Kenneth W. Simons, *Rethinking Mental States*, 72 B.U. L. REV. 463, 464 (1992). Simons also explicitly counts strict liability as a "Conduct" state that is "not a true mental state." *Id.* at 465 tbl. 2.

207. MODEL PENAL CODE § 2.02(2)(d) (Official Draft and Revised Comments 1985).

mental state either: "A crime that does not require a *mens rea* element."[208] Here, strict liability is not even a sort of *mens rea* that does not require any psychological states.

The Manifestation of *Mens Rea* Rationale can handle negligence and strict-liability omissions because the *mens rea* might still *cause* the omission and thereby become *manifest* in the omission. But the *causing* and the *manifestation* would have nothing in particular to do with the defendant's actual psychology or the bearing that her psychology had, if any, on her control or moral responsibility. A state of negligent disregard might in some sense *cause* the defendant's failure to perform a legal duty. And a state of, perhaps non-negligent, disregard also might in some sense *cause* her failure to perform a legal duty. In negligence and strict liability omission cases, we might say that the *mens rea causes*, and becomes *manifest* in, the omission by constituting, part or all of, the omission. To offer an example for analogical purposes: A's sister's having a baby caused A to become an aunt. The sister's having a baby constitutes A's becoming an aunt—the cause and effect are not logically independent. One might also, loosely, say that the sister's having a baby is manifested in A's becoming an aunt.

But for the agent-causal retributivist, the mere manifestation of *mens rea* in this *causal-logical* sense is not the crucial justificatory point. For a retributivist, the crucial point is always whether the defendant deserves punishment. Since the defendant's actual psychology and agent-causal control are irrelevant in negligence and strict liability omission cases, it would seem that for the agent-causal retributivist to justify the VAR in such cases, she must explain how a defendant could deserve punishment without relying on or referring to the defendant's actual psychological states or the degree of the defendant's agent-causal control.

An agent-causal retributivist could offer such an explanation, however. A defendant could deserve punishment for an omission if it seems justifiable for her to be subject to punitive liability independently of whether she, as an agent, *tries*, in the technical sense, to do, or to not-do, anything. Only certain omissions are ever included in the *actus reus* of a crime.

208. *Strict-Liability Crime*, BLACK'S LAW DICTIONARY 378 (7th ed. 1999).

And those omissions are failures to perform legally imposed duties. If in all cases in which the law imposes such duties it seems plausible that failing in the duty would ground moral responsibility and desert, then a retributivist, whether an agent causalist or not, would be able to justify including some omissions within the VAR. (Although the retributivist might complain that it is misleading to include omissions within a principle called the "Voluntary Act Requirement." Maybe it would be better to enact a separate "Limited Class of Omissions Requirement" and then insist that the fundamental predicate of criminal liability would be satisfying either the VAR or the separate "omissions" requirement.)

In *Manon*, the defendant was the infant's mother.[209] She had not given him up for adoption, and she had not given up her parental rights.[210] It may seem plausible that she deserves punishment for neglecting her son even if she did not *try* to neglect him. To offer another example, in *Commonwealth v. Levesque*,[211] the court reversed a dismissal of the defendants' indictment for manslaughter. The defendants were alleged to have (a) accidentally started a warehouse fire and (b) failed to report the fire.[212] Eventually, six fire fighters perished in the blaze.[213] The defendants' indictment was dismissed because the defendants successfully argued that they "had no legal duty to report the fire and [that] their failure to act did not satisfy the standard of wanton and reckless conduct required for manslaughter charges."[214] The *Levesque* court reversed the dismissal, however:

> It is true that, in general, one does not have a duty to take affirmative action[;] however, a duty to prevent harm to others arises when one creates a dangerous situation, whether that situation was created intentionally or negligently.[215]

209. People v. Manon, 640 N.Y.S.2d 318, 319 (App. Div. 1996).
210. *See generally id.*
211. Commonwealth v. Levesque, 766 N.E.2d 50 (Mass. 2002).
212. *Id.* at 53.
213. *Id.*
214. *Id.*
215. *Id.* at 56.

> Where a defendant's failure to exercise reasonable care to prevent the risk he created is reckless and results in death, the defendant can be convicted of involuntary manslaughter. Public policy requires that "one who creates, by his own conduct . . . a grave risk of death or injury to others has a duty and obligation to alleviate the danger."[216]

It may seem plausible that the *Levesque* defendants deserve punishment for not taking steps to prevent the risk that they created even if they did not *try* not-to-prevent it. Generally, as long as the legal duty imposed is similar to those imposed in *Manon* and *Levesque*, the agent-causal retributivist can explain how a defendant could, by omission, satisfy the VAR and deserve punishment. And this could be done without relying on or referring to the defendant's actual psychological states[217] or the degree of the defendant's agent-causal control.

C. *Habitual Action Cases*

In the third type of case, there is criminal liability even though the defendant acts out of habit. Habitual actions count as voluntary acts for the purposes of the VAR. As MPC § 2.02(d) clarifies, not counting as a voluntary act for the purposes of the VAR is "a bodily movement that . . . is not a product of the effort or determination of the actor, either conscious or habitual." This suggests that a habitual bodily movement is a product of the effort or determination of the actor and is therefore a voluntary act.

Yaffe devises a hypothetical:

216. *Id.* at 57 (citations omitted).

217. In the case of a reckless omission some reference to the defendant's actual psychology would be necessary because "[a] person acts recklessly with respect to a material element of an offense when he *consciously disregards* a substantial and unjustifiable risk that the material element exists or will result from his conduct . . . ." MODEL PENAL CODE § 2.02(2)(c) (Official Draft and Explanatory Notes 1985) (emphasis added). But the crucial point remains that the defendant's agent-causal power would be irrelevant.

> Consider a defendant who has been trained by the military to spin around and fire immediately, and without thinking, on a threat behind him. This behavior has become, thanks to his training, habitual. Is he to be held guilty of a crime when, at the local firing range, he spins and fires on a person behind him who yells something threatening? . . . The bodily movements in cases such as this are routinely taken to provide an acceptable basis for criminal liability.[218]

Yaffe points out that in such a case, a court would assume that the defendant willed the habitual behavior, even if the court was unsure whether the behavior was willed. As Yaffe explains:

> Normally, our ignorance about a feature of pertinence to criminal liability is enough to supply reasonable doubt, and thus enough to support an acquittal. But not in habitual action cases. If there is reasonable doubt about whether the defendant's bodily movement was voluntary deriving from the fact that it is habitual, that reasonable doubt fails to undermine the case for guilt. We treat habitual bodily movements . . . as though they were voluntary acts in the legal sense, even though we have no idea whether they are in fact.[219]

Thus, habitual action cases depart from the fundamental legal principle that requisites for criminal liability must be proven beyond a reasonable doubt before punishment is imposed. Even if it seemed ludicrous to suggest that a defendant acted voluntarily in any usual sense of the term *voluntarily*, acting on the basis of habit would dispositively establish voluntary action for legal purposes. And as Yaffe further clarifies, courts will also consider the VAR satisfied in habitual action cases without regard to whether the defendant was at

218. Yaffe, *supra* note 2, at 177.
219. *Id.*

fault in being in a situation where his habit might be "triggered" and cause harmful results:

> The defendant in [the firing range] case . . . need not be shown to have been at fault for being in the circumstances in which he found himself, nor would the behaviors that got him there need to be shown to have been voluntary. The mere fact . . . that the relevant bodily movements were the product of habit is sufficient to show there to be compliance with the VAR in assigning a guilty verdict.[220]

To modify the firing range case to challenge the idea that habit should always automatically establish voluntariness, imagine that the conditioning that ingrained the soldier's shooting habit was intense and irresistible. We could imagine a process similar to the conditioning not to commit harmful acts imposed on Alex DeLarge in Stanley Kubrik's movie, "A Clockwork Orange."[221] The soldier volunteers to be subjected to this "rewiring" to empower him to complete vital military missions that save thousands of lives. Upon retirement from the military, knowing of the potential danger his habit poses, he never goes near firearms. He assiduously avoids any situation in which his latent automatic response might be triggered.

One day, some evil masterminds decide to kill their nemesis, Victor, in an elaborate way. So they drug and kidnap both Victor and the habituated soldier, taking both to the local firing range. They put a loaded gun in the soldier's hand, and they place Victor behind the soldier. Very close to Victor's head, they position a speaker ready to project a pre-recorded sound that the masterminds know will *trigger* the soldier's ingrained, automatic habit. Immediately after the drugs wear off and both kidnap victims fully come to, the masterminds play the

220. *Id.* at 183.

221. In this movie, Alex commits anti-social acts. He is caught and eventually released back into society, but only after having undergone an intense regimen of conditioning involving drugs and other invasive techniques, which renders him violently ill whenever he tries to commit further anti-social acts.

recording remotely. The rest of the hypothetical is the same as Yaffe's. The soldier spins and fires, killing Victor. If the soldier is tried for homicide, then the court will rule that the VAR is satisfied. This is not to say that the soldier ultimately will be found guilty. But if he is acquitted, it will not be because the VAR was not satisfied.

How might an agent-causal retributivist justification of the VAR handle such a case? The agent causalist faces a significant difficulty here. The habit of the militarily-trained defendant was apparently ingrained to the point where he lacked control over spinning and firing. The most straightforward agent causalist assessment would be that the VAR should not be considered satisfied because the defendant did not cause his conduct. Instead, his training took over and made him do what did in an event-causal, deterministic manner. Essentially, the masterminds used the soldier as a weapon. The soldier was no more responsible, morally, than the gun was. He therefore lacked desert:

> Under one construal, habitual actions are simply "triggered" by perception of the environment. It is because he hears something threatening behind him that the soldier . . . spins and fires. But given that it is not in his control that he should hear the threat, it is not under his control that he should spin and fire on the person who issued it.[222]

It is tempting to argue that, at least usually, there is no problem treating habitual actions as voluntary. For example, imagine that Fred repeatedly drives his car quickly through a fairly remote intersection to save time.[223] He goes to work at 2:00 AM, and for years there has never been anyone else on the road then. Fred's habit regarding that intersection has become virtually automatic. Then one night someone crosses that intersection on foot a little after 2:00 AM. Fred notices her, but only a couple of moments before getting to the intersection.

222. Yaffe, *supra* note 2, at 181.

223. This example is a modified version of an example in Yaffe, *supra* note 2, at 174.

Usually, he would be able to stop in time, but Fred's habit kicks in, and the pedestrian is run over. If Fred is charged with a crime, then presumably he will not be able successfully to defend by arguing that the VAR was not satisfied. And even in an extreme, contrived scenario such as the modified firing range case, some other principle of criminal law would prevent the soldier from being found guilty.[224] From a pragmatic point of view, it is acceptable to pretend that the soldier voluntarily spun and fired at Victor as long as, in the end, the soldier is not found guilty of a crime.

Insofar as we care about punishing persons only when it is the right thing to do, this pragmatic argument does reasonably well. By treating habitual actions as *per se* voluntary, we get things right, vis-à-vis voluntariness, most of the time, and we never punish someone who does not voluntarily act. So for practical purposes, we can defend the law's deviation from requiring proof beyond a reasonable doubt of any factor pertinent to criminal liability.

But this approach is *second-best*. When applying and justifying a doctrine as fundamental as the VAR, it should matter not just that we wind up getting the right answer as to guilt or innocence. It should also matter that we get the right answer for the right reason. At stake is justifying not just coercive governmental intrusion into the lives of criminals that causes them suffering. Punishment hurts criminals on purpose. Ideally, we should have a good justificatory reason behind every aspect of doing *that*. And treating the soldier as if he acted voluntarily would be based on a falsehood.[225]

Consider Yaffe's account of how the Manifestation of *Mens Rea* Rationale handles habitual action cases in general:

224. Perhaps the soldier would not satisfy the requisite *mens rea*.

225. But consider: There is a significant cost savings in habitual action cases if no proof is required that the habit involved leaves the defendant sufficiently in control. Requiring proof of other elements of *mens rea* and *actus reus* is already expensive, and those requirements afford the defendant significant protection against wrongful conviction. (I wish to thank David A. Simon for suggesting this in conversation.)

This might justify the law's practice from an economic point of view. But such a justification would compromise the main idea behind the VAR—assuring control sufficient for criminal liability. We might ultimately need to decide whether respecting the main idea behind the VAR in all cases is worth it. If it is not, then the economic justification might just have to do.

> The rationale for treating convictions for habitual actions as complying with the VAR is . . . [that] some habitual actions can be manifestations of objectionable mental states in the sense that matters to morality, even if they are not guided by one's conscious volitions. One benefit of a habit is that it produces conduct for which there are reasons without the agent taking the time to reflect on and weigh those reasons . . . . However, a byproduct of this valuable feature of habits is that they override our tendencies to withhold action in the face of reasons to do so.[226]

Yaffe argues that habits often connect a *mens rea* with an *actus reus* to establish criminal liability. Fred might be considered reckless when he strikes the pedestrian with his car. His habit of running through the intersection connects his recklessness to the striking of the pedestrian such that Fred's recklessness becomes manifest in the striking. If the law did not treat Fred's habitual action as voluntary, then the VAR's function of assuring correspondence between *mens rea* and *actus reus* would be undermined.

This argument works well in many cases, such as Fred's. When it is Fred's recklessness that needs to get connected to the consequences of his conduct, things seem to turn out right vis-à-vis the VAR. But as Yaffe also points out, a court will treat a defendant's habitual action as voluntary without any regard to whether the defendant is responsible for being in a situation in which his habit might get triggered with disastrous results. Fred was reckless. He consciously disregarded the risk posed by heading toward the intersection, given is habit of going right through it without stopping. Fred was at fault for putting himself into a dangerous *triggering* situation. And this is why the Manifestation of *Mens Rea* Rationale works so well in Fred's case.

But what about the modified firing range case? The soldier was entirely free from fault for being in a situation in which his

226. Yaffe, *supra* note 2, at 188.

habit was dangerously triggered. And the soldier had a good, perhaps even heroic, reason to get habituated to turn and shoot in the first place. Unlike Fred's case, there is no recklessness to get connected to the consequences of the soldier's conduct to assure that his conduct manifests recklessness. And it is difficult to see what *mens rea* there may be other than recklessness to connect to and thereby be made manifest in the soldier's conduct. There seems to be no point, in the modified firing range case, to assuring the manifestation of *mens rea* to justify imposing criminal liability upon the soldier.

And this problem carries over generally to cases where a defendant is not morally blameworthy for being in a situation in which his habit might threaten the interests that criminal law is designed to protect. In sum, it seems that in cases such as the modified firing range case it is not defensible to treat the defendant as though he acted voluntarily. And the agent-causal retributivist has a straightforward explanation for this. It is not defensible because the VAR is supposed to assure that the defendant has control, and in such cases the defendant lacks control because he does not agent-cause his conduct.

VII. Conclusion

For retributive justifications of the VAR to be plausible, agent causalism must be true. The ALI clarifies that the main idea behind the VAR is to assure that a defendant controls any conduct for which she is subject to criminal liability. For a retributivist, control is important because without it, the defendant cannot be morally responsible for her conduct and therefore cannot deserve punishment. And for the defendant to have the requisite control, she must agent-cause her conduct. This result is significant because agent causalism is contentious. If agent causalism is false, then retributivism—one of the two major theories of punishment in the legal academic literature and judicial opinions—cannot contribute to justifying the fundamental predicate of criminal responsibility.

The significance of this result is enhanced when one considers that agent causalism is not incoherent and is at least plausible. It should be taken seriously as a possible solution to the problem of free will. Additionally, if agent causalism is true,

then the retributivist can account not only for why the VAR is justified in paradigmatic cases of criminal conduct. The agent-causal retributivist can handle unusual cases too—including cases featuring complex unconscious behavior and crimes of omission. Finally, the agent-causal retributivist can accommodate most cases in which criminal conduct is caused directly by habit. It is consistent with agent-causal retributivism to rely on the Manifestation of *Mens Rea* Rationale for those cases, since punitive desert requires correspondence between *mens rea* and *actus reus*. And in cases such as the modified firing range case, the agent-causal retributivist can explain why the law is not justified in failing to require proof beyond a reasonable doubt of a factor necessary for criminal liability. At least in such cases, the law should not continue to assume that a defendant's bodily movement is voluntary simply because it is habitual.