Tulsa Law Review

Volume 54 | Issue 3 Article 6

Spring 2019

Retributivism, Ultimate Responsibility, and Agent Causalism

Christopher P. Taggart

RETRIBUTIVISM, ULTIMATE RESPONSIBILITY, AND AGENT CAUSALISM

Christopher P. Taggart[1]

Except for limited forms of omissions liability, Anglo-American criminal law generally requires a criminal defendant, D, to perform a voluntary action before imposing criminal liability. Further, D must be morally responsible for performing the action for D to deserve punishment for doing it. So, a puzzle about moral responsibility connected to longstanding debates about determinism and free will, a puzzle that implies that D is never morally responsible for performing any action, must have a moral-responsibility-preserving solution for any form of retributivism to be true. One compatibilist solution denies that moral responsibility requires what has been termed "ultimate responsibility." Whether ultimate responsibility is required for moral responsibility is a contested issue. And, if ultimate responsibility is required for moral responsibility, then the compatibilist solution is unavailable. This article argues that, if ultimate responsibility is required for moral responsibility, then, unless both indeterminism and agent causalism are true, any form of retributivism is false.

1. Lecturer in Law, University of Surrey School of Law. I wish to thank members of the Surrey Centre for Law and Philosophy for helpful comments on earlier drafts.

I. INTRODUCTION

Two longstanding sets of philosophical debates are significantly related. The first concerns how, if at all, retributivism is a defensible element of a theory of the justification of legal punishment. The second concerns how best to resolve various puzzles surrounding causal determinism and free will. A very broad question arises: What position should one take regarding one set of debates given the position one takes regarding the other set? This question delineates the general topic of this article.

More specifically, this article argues that, on the assumption that a criminal defendant's (D's) moral responsibility for committing a crime requires D's "ultimate responsibility" for committing that crime, any retributive theory of punishment is true only if both indeterminism and agent causalism are true. In other words, on the assumption that ultimate responsibility is required for moral responsibility: (1) If determinism is true, then no retributive theory of punishment is true; and, (2) if agent causalism is false, then no retributive theory of punishment is true, even if determinism is false. This paper, therefore, presents no arguments against retributivism backed by a compatibilist position that denies the necessity of D's ultimate responsibility for committing a crime for D to deserve punishment for committing that crime.[2]

To defend its thesis, the article proceeds as follows. Part II explains "retributive reasons" and argues that such facts must exist for any form of retributivism to be true. Section III presents a puzzle related to philosophical debates concerning determinism, free will, and the control that agents exercise when they perform voluntary actions. Part IV discusses "ultimate" responsibility for doing something and explains why, *if* "true" moral responsibility for an action requires ultimate responsibility for it, *then* determinism forecloses any retributivism-preserving solution to Part III's puzzle. Part V introduces agent causalism and argues that, even if determinism is false, agent causalism must be true for there to be a solution to Part III's puzzle that preserves the possibility of retributivism. Part VI concludes by briefly recapitulating the main line of argument.

II. RETRIBUTIVE REASONS

Retributive theories of punishment[3] emphasize a criminal defendant's, D's,[4] negative moral desert.[5] The retributivist idea of moral desert is familiar: when D does

---

2. Although I think that compatibilism is false, I offer no arguments for incompatibilism in this paper.

3. By "punishment," this paper means punishment imposed under criminal law. Theories of punishment answer at least two questions: "To whom may punishment be applied? How severely may we punish?" H.L.A. HART, *Prolegomenon to the Principles of Punishment*, *in* PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 1, 2 (1968). If a punishment of D is justified, then it is morally permissible for the state to inflict that punishment upon D. This article presupposes "negative" retributivism's view of desert as necessary to justify the severity of a punishment of D. Punishing D more severely than D's desert allows is morally impermissible.

4. It is assumed throughout that D is a generally competent adult of at least average intelligence.

5. Sometimes, desert is not *moral* desert. "People make desert claims in a wide variety of strikingly different contexts. We might talk, for example, about one person who deserves a promotion, someone else who deserves a good grade, and some third person who deserves to win the race they are competing in." SHELLY KAGAN, THE GEOMETRY OF DESERT 4 (2012). Also, "negative moral desert" is desert that calls for a bad desert object, such as punishment, instead of a good one, such as praise.

something morally wrong, D deserves punishment for doing so, and the morally worse the wrongdoing, the more severe the punishment D deserves for doing it. Retributivism places justificatory weight on D's desert when addressing how severely to punish D for something that D has done.

A *retributive reason* is a fact about what punishment, if any, D deserves for doing something.[6] For example, if D is charged with an offense, C, that D did not commit, then D deserves no punishment for C. That D is innocent of C (and thus deserves no punishment for committing C) is a retributive reason not to punish D for committing C. On any form of retributivism, retributive reasons are (at least) *relevant* to whether punishing D is justified.[7] And, retributive reasons have a characteristic form: "[D] deserves X in virtue of F, "where [D] is a person, X is a mode of treatment, and F [is] some fact about [D]."[8] If X is legally-imposed hard treatment, death, censure, or the like, then F must be a fact about something D voluntarily did.[9]

Placing justificatory weight on retributive reasons is consistent with different forms of retributivism, some of which are very modest. According to some, "just punishment [is] one good among many, and one that can be outweighed by other goods that punishing the deserving puts at risk."[10] Retributivists could even think that criminal law's *principal* overall objective is preventing social harm, instead of assuring that people get (or get no more than) their just deserts.[11] Regardless of how modest the form of retributivism, the desert subject must be a moral agent, D, who performs an action that is the desert basis. Only a moral agent could deserve or not deserve punishment. And the action that is the desert basis can be referred to as a "wrongdoing."[12] Finally, if there are no retributive facts, then no form of retributivism, however modest, is true.

---

6. This definition of "retributive reason" is stipulative. So, I am not *assuming* that there are any retributive reasons. The definition clarifies what such a thing would be if there were any. Sometimes, I will interchangeably use the term "retributive fact."

7. Arguably, "[a] legal philosopher does not [even] qualify as a retributivist if he . . . awards [retributivism] only a peripheral role in his rationale for . . . punitive sanctions." Douglas Husak*, Broad Culpability and the Retributivist Dream*, 9 OHIO ST. J. CRIM. L. 449, 450 (2012).

8. JOEL FEINBERG, *Justice and Personal Desert*, *in* DOING & DESERVING: ESSAYS IN THE THEORY OF RESPONSIBILITY 55, 61 (1970).

9. This article does not adopt any position about what, specifically, D should get when D gets what D deserves. This article brackets omissions liability and focuses on criminal liability for voluntary action.

10. L. ALEXANDER & K. FERZAN, CRIME AND CULPABILITY 8–9 (2009). Along similar lines, "[n]early every retributivist finds room for consequentialist considerations . . . somewhere in his account. . . . [T]o count as retributive . . . [a theory] need only regard desert and blame as central to attempts to provide answers to normative questions about . . . punishment." Husak, *supra* note 7, at 450 n.4.

11. *See* ALEXANDER & FERZAN, *supra* note 10, at 4. (arguing that "the criminal law's primary concern is the prevention of harm" while endorsing a moderate form of retributivism). "[I]t is perfectly consistent to assert *both* that the General Justifying Aim of . . . punishment is its beneficial consequences *and* that the pursuit of the General Aim should be qualified or restricted out of deference to [retributive] principles of Distribution." HART, *supra* note 3 at 5.

12. By "wrongdoing," I mean an action that is *morally* or *ethically* bad or wrong, remaining neutral about what the moral badness/wrongness of a morally bad/wrong action amounts to. This paper considers punishment for *malum in se* offences only. In different contexts, one could refer to the commission of a *malum prohibitum* offence as a "wrongdoing" or distinguish "criminally illegal" wrongs that are also moral wrongs from those that are wrongs only because criminally illegal. I am ignoring such complications.

III. RETRIBUTIVISM, CONTROL, AND A PUZZLE ABOUT MORAL RESPONSIBILITY

For a retributive fact about D to exist, D must deserve punishment for some wrongdoing that D performs. Such desert requires D to be morally responsible for that wrongdoing. I will describe D's wrongdoings as blameworthy and D, the agent, as morally responsible for performing it. D's wrongdoing is blameworthy only if D is morally responsible for performing it.[13] Further:

> When we hold [D] [morally] responsible for acting wrongly, [D] is not the same as the feature of [D's] act for which we hold [D] responsible. Because our blame and punishment are directed at the agent [D] but are justified (if they are) by the wrong-making features of what [D] has done, their grounding must include some appropriate relation between the agent [D] and [D's] act's wrong-making features.[14]

Arguably, for D to be related in the right way to the wrong-making features of D's blameworthy actions two conditions must be satisfied: (1) D must have sufficient control over what D does, and (2) D must have sufficient knowledge about what D does, including the context in which D acts. We can refer to (1) as a *voluntariness condition* and (2) as an *epistemic condition*.[15] I will assume that, for a retributive justificatory "grounding" of D's punishment to be adequate, the relation between D and D's act's wrong-making features must accommodate *both* the voluntariness condition and the epistemic condition. The focus going forward, however, will be the voluntariness condition.[16]

Retributive facts about D exist only if D performs blameworthy actions. Restricting our attention to D's wrongdoings that are criminal offences, as a first approximation, D's action is blameworthy if D "commit[s] the *actus reus* of [the] offense with [any] morally blameworthy state of mind."[17] However, "[m]any aspects of blame are not matters of *mens rea*. . . . [Some] . . . involve *actus reus*."[18] The blameworthiness of D's wrongdoing depends on factors other than the sorts of mental states that D had, or should have had, when D acted. The "conduct" element of any *actus reus* features a voluntary act. This "voluntary act requirement" (VAR) requires that D have the right sort of *control* over what D does when D does it.[19] And, insofar as the VAR is to be justified along retributivist lines,

---

13. Moral responsibility is a necessary condition for praiseworthy actions as well. D can be morally responsible for doing something that is not blameworthy. But, if D's action is blameworthy, then D is morally responsible for doing it.

14. GEORGE SHER, WHO KNEW? 147 (2009).

15. *See id.* (employing this terminology). This dual requirement for moral responsibility has been expressed in different ways: "Acts that are voluntary receive praise and blame, whereas those that are involuntary receive pardon and sometimes pity too. . . . Actions are regarded as involuntary when they are performed under compulsion or through ignorance." *Id.* at 3 (quoting ARISTOTLE, THE ETHICS OF ARISTOTLE: THE NICOMACHEAN ETHICS 111 (J.A.K. Thompson, trans. 1955)). "[A]ny being who is held responsible must be sufficiently rational and autonomous to be a moral agent," for "only such beings are capable of being morally culpable." MICHAEL S. MOORE, PLACING BLAME 403 (1997).

16. In the remainder of this article, for any action discussed for which an agent might be morally responsible, I will *assume* that the epistemic condition is satisfied. Going forward, moral responsibility turns on whether the voluntariness condition is satisfied.

17. JOSHUA DRESSLER, UNDERSTANDING CRIMINAL LAW 118 (6th ed. 2012) (emphasis added).

18. Husak, *supra* note 7, at 458.

19. The American Model Penal Code formulates the VAR at § 2.01(1): "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

for D to *deserve* punishment for doing something, D must exercise the right kind of *control* when D does it.

In sum: (a) retributivism requires the existence of retributive facts about D; (b) the existence of retributive facts about D requires D's punitive desert; (c) D's punitive desert requires D's blameworthy conduct; and (d) D's blameworthy conduct requires D to exercise the "right sort" of control when D acts. Thus, serious doubts about whether D has the "right sort" of control over D's actions threaten retributivism.

The longstanding debates about determinism and free will have raised such doubts, and so, the next section will engage some of those debates. To set the stage, this section concludes with a major assumption, some brief definitions, and a puzzle.

First, the assumption: "[D]'s [voluntary] actions are those events involving [D] caused (in the right way) by" D[20] and D's intentional states.[21] To get an idea of what causation "in the right way" is, consider a "deviant" causal chain between an agent's intentional states and his body's behaviour:

> A 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.[22]

The climber's intentional states do not fit into the right sort of causal explanation of the loosening of his hold to make the hold-loosening one of the climber's voluntary actions. The causal explanation of the climber's body's behaviour is not an *intentional explanation*. "Intentional explanations explain a bit of behavior . . . by making it reasonable in the light of certain beliefs, intentions, [and] desires . . . [of] the agent."[23] If, instead, the climber voluntarily let go to save himself, then the hold-loosening would have an accurate intentional explanation, based on the "right sort" of causation. In that case, the climber's intentional states would causally affect the climber's body's behaviour in a manner that made the hold-loosening reasonable in virtue of the contents of the causally operative intentional states.[24]

---

he is physically capable." MODEL PENAL CODE § 2.01(1) (AM. LAW INST. 2018). The American Law Institute clarifies that "the term 'voluntary' as used in [§ 2.01(1)] . . . focuses upon conduct that is within the *control* of the actor. " *Id*. cmt. 1 (emphasis added).

20. *If* there were no defensible distinction between D and D's intentional states operating as causes, *then* this assumption would pertain only to causation "in the right way" by D's intentional states. I will consider the distinction between D and D's intentional states operating as causes in Part V.

21. Peter A. Graham, *The Standard Argument for Blame Incompatibilism*, 42 NOÛS 697, 703 (2008).

22. DONALD DAVIDSON, *Freedom to Act*, *in* ESSAYS ON ACTIONS AND EVENTS 63, 79 (2d ed., Oxford Univ. Press 1980).

23. DANIEL C. DENNETT, *Mechanism and Responsibility*, *in* BRAINSTORMS: PHILOSOPHICAL ESSAYS ON MIND AND PSYCHOLOGY 233, 236 (1981). Dennett's original definition of an intentional explanation does not refer to an agent's intentional states but instead refers to intentional states that are "ascribed to" an agent, such as D, by an analyst, A. For Dennett, A's ascription of intentional states to a "system," such as D, is part of A's adoption of a useful heuristic strategy toward D. I am assuming that, when D's intentional states play a causal role (in the sense of *efficient causation*) that makes an intentional explanation of D's behavior true, D *really has* those intentional states, *independently* of the heuristic strategy that A chooses for A's systematic, predictive purposes. I am assuming that A's heuristic strategy is useful *because* D *really has* those intentional states.

24. It bears reemphasis that all of this is part of a major *assumption* about what a voluntary action is. Defending this contested assumption exceeds this paper's scope. To clarify one point, it is *not* part of the

Turning to the definitions: *Free will* is "the unique ability of persons to exercise control over their conduct in the manner necessary for moral responsibility."[25] "*Determinism* is the thesis that the past and the laws of nature together determine, at every moment, a unique future. . . ."[26] Indeterminism "is the denial of determinism."[27] *Compatibilism* is the thesis that it is possible both that determinism be true and that D have free will.[28] *Incompatibilism* is the denial of compatibilism.[29]

With the assumption and definitions on the table, consider a puzzle about moral responsibility (PMR): Moral responsibility for blameworthy action seems to be incompatible both with determinism and indeterminism.[30] Moral responsibility for blameworthy action seems, therefore, to be impossible. But sometimes agents seem to be morally responsible for performing blameworthy actions. The impossible therefore seems to exist.[31]

A solution to PMR would resolve the seeming contradiction. And, for such a solution to accommodate retributive facts, the solution must accommodate D's having the "right sort" of *control* needed to be morally responsible for performing a blameworthy wrongdoing.

IV. DETERMINISM, CONTROL, AND "ULTIMATE" MORAL RESPONSIBILITY

For PMR to threaten retributivism, there must be *prima facie* reasons to think that the sort of control necessary for D's moral responsibility for blameworthy action is incompatible with *both* determinism and indeterminism, as per hard incompatibilism. On one conception of the control that D must have to be morally responsible for committing

---

assumption that *reasons themselves* stand in causal relations to states and events. The things that stand in such relations under the assumption are intentional states, whose contents are reasons.

25. Michael McKenna & D. Justin Coates, *Compatibilism*, STAN. ENCYCLOPEDIA OF PHIL. (last modified Feb. 25, 2015), https://plato.stanford.edu/entries/compatibilism/. I am adopting this definition by stipulation. So, I am not *assuming* that D has this ability. The definition, as far as it goes, states an ability that D *would* have *if* D had free will.

26. Peter van Inwagen, *How to Think about the Problem of Free Will*, 12 J. ETHICS 327, 330 (2008). More precisely, determinism is true if and only if the set of the actual laws of nature is *deterministic*, where: "[a] set of laws of nature is deterministic just in case there is one, and only one, distinct possible world for each initial state of the world compossible with that set of laws. . . . So, if the [set of laws of nature of the] actual world is deterministic and [D] does not φ at t, any world in which [D] does φ at t must either have different laws of nature or a different initial state (or both) from those of the actual world. " Graham, *supra* note 21, at 701.

27. Inwagen, *supra* note 26, at 330.

28. The possibility here is metaphysical: Compatibilism is the thesis that there are possible worlds in which determinism is true and D has free will. Compatibilists could consistently deny both that determinism is true and that D has free will. *Soft determinists* are compatibilists who think that determinism is true and D has free will.

29. Since incompatibilists think that there is no possible world in which determinism is true and D has free will, incompatibilists would deny the soft determinist's claim that both are true in the actual world. A *hard incompatibilist* is an incompatibilist who thinks that "there is no free will if determinism is false." McKenna & Coates, *supra* note 25. In other words, a hard incompatibilist thinks that free will is consistent with *neither* determinism *nor* indeterminism and is, therefore, metaphysically impossible.

30. Note that I am not endorsing every step of PMR as compelling. PMR serves to frame subsequent discussion.

31. This formulation of PMR apes Peter van Inwagen's formulation of the "Problem of Free Will": "[f]ree 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." Peter van Inwagen, *Free Will Remains a Mystery: The Eighth Philosophical Perspectives Lecture*, 14 PHIL. PERSP. 1, 11 (2000).

a crime, C, at t, D must, at t, be able not to commit C. The Principle of Alternate Possibilities captures this conception: "[A] person is morally responsible for what he has done only if he could have done otherwise."[32]

Harry Frankfurt poses a counterexample to this principle:

> Jones has resolved to shoot Smith. Black has learned of Jones's plan and wants Jones to shoot Smith. But Black would prefer that Jones shoot Smith on his own. However, concerned that Jones might waver in his resolve to shoot Smith, Black secretly arranges things so that, if Jones should show any sign at all that he will not shoot Smith (something Black has the resources to detect), Black will be able to manipulate Jones in such a way that Jones will shoot Smith. As things transpire, Jones follows through with his plans and shoots Smith for his own reasons. No one else in any way threatened or coerced Jones, offered Jones a bribe, or even suggested that he shoot Smith. Jones shot Smith under his own steam. Black never intervened.[33]

In the counterexample, Jones seems morally responsible for shooting Smith, even though Jones could not have done otherwise when he pulls the trigger.

As John Martin Fischer and Mark Ravizza might say, Jones has "guidance control" but lacks "regulative control":

> [S]uppose that Sally is driving her car. . . . Sally wishes to make a right turn. As a result of her intention to turn right, she . . . carefully guides the car to the right. Further . . . assume that Sally was able to form the intention not to turn the car to the right but to turn the car to the left instead . . . [and] that, had she formed such an intention, she would have turned the steering wheel to the left and the car would have gone to the left. . . . Sally guides the car to the right, but she could have guided it to the left. She controls the car, and . . . she has a certain sort of control over the car's movements. Insofar as Sally [] guides the car in a certain way, she [exercises] "guidance control." Further, insofar as Sally also has the power to guide the car in a different way . . . she has "regulative control."[34]

Returning to Frankfurt's counterexample, since Jones lacks regulative control, possesses guidance control, and seems morally responsible for shooting Smith, the sort of control necessary for moral responsibility appears to be some form of guidance control. Whether this appearance is correct turns, in part, on what being "truly" morally responsible for a criminal wrongdoing requires. On one view, the sort of "true" moral responsibility necessary for punitive desert is *ultimate responsibility*. D is ultimately responsible for a wrongdoing if D "[is], or [is] responsible for, the ultimate (determining) causes of [D's wrongdoing], having personally determined the (determining) causes of [D's] actions 'all the way back."[35]

The "Consequence Argument," which shows that, *if determinism is true*, then D is not ultimately responsible for any wrongdoings, helps clarify what "all the way back"

---

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

33. McKenna & Coates, *supra* note 25 (summarizing Frankfurt's counterexample from *Alternate Possibilities and Moral Responsibility*, 66 J. PHIL. 829 (1969)).

34. JOHN MARTIN FISCHER & MARK RAVIZZA, RESPONSIBILITY AND CONTROL: A THEORY OF MORAL RESPONSIBILITY 30–31 (1998).

35. K.E. BOXER, RETHINKING RESPONSIBILITY 14 (2013). I am stipulatively adopting Boxer's definition. Therefore, I am not *assuming* that D is ultimately responsible for anything.

amounts to. First, a pithy version of the Consequence Argument:

> If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us.[36]

A recent formalisation of the Consequence Argument by Brian Cutter more perspicuously unpacks the argument's steps.[37] Cutter's formalisation requires some apparatus. In addition to the standard inference rules of modal and propositional logic, we need, first, an operator, "N": $Np =_{df} p$, and every agent S is such that, for anything S can do, if S were to do it, it would still be the case that p. Second, we need two inference rules involving N, where "$\Box$" is the standard modal operator for broadly logical necessity:

(α) $\Box p \vdash Np$
(β) $N(p \supset q), Np \vdash Nq$.[38]

Third, let P_0 be a complete specification of the world at some moment, t_0, before the existence of any human beings. Fourth, let L be the conjunction of all the laws of nature. And fifth, let P be a true proposition about some arbitrary thing that happens after t_0.

(1) $\Box((P_0 \,\&\, L) \supset P)$	[premise – determinism]
(2) $\Box(P_0 \supset (L \supset P))$	[from 1, standard modal and propositional logic]
(3) $N(P_0 \supset (L \supset P))$	[from 2, (α)]
(4) NP_0	[premise]
(5) $N(L \supset P)$	[from 3, 4, (β)]
(6) NL	[premise]
(7) NP	[from 5, 6, (β)]

Letting P be the proposition that D commits C at t (obviously, later than t_0), the Consequence Argument shows that, for anything D can do, if D were to do it, it would still be the case that D commits C at t. "All the way back" in the definition of "ultimate responsibility" can be understood in reference to t_0. The Consequence Argument shows that D cannot personally determine the (determining) causes of (D's) actions because those determining causes are "fully operative" at t_0, before D, personally, could cause or determine anything.

Thus, *if determinism is true*, then, D is not ultimately responsible for committing C. So, *if* moral responsibility for committing C requires ultimate responsibility for committing C, *then* determinism forecloses D's moral responsibility for committing C, *even if* D exercises guidance control in committing C. Hence, *if* ultimate responsibility is required for "true" moral responsibility, *then* determinism forecloses the existence of any retributive facts about D. This raises a crucial question: Does "true" moral responsibility require ultimate moral responsibility?

A compatibilist might argue that ultimate responsibility is *not* required for moral responsibility. In Frankfurt's counterexample, Jones seems to be morally responsible for

---

36. PETER VAN INWAGEN, AN ESSAY ON FREE WILL v (1983).

37. *See* Brian Cutter, *What is the Consequence Argument an Argument for?*, 77 ANALYSIS 278, 280 (2017). Cutter's formalisation is based on a formalisation in VAN INWAGEN, *supra* note 36.

38. *See* VAN INWAGEN, *supra* note 36, at 94.

shooting Smith when Black does not intervene. Jones's shooting Smith is the output of a moral-responsibility-conferring psychological mechanism that is Jones's alone—a mechanism that operates unencumbered by "outside" or "external" interference. If Jones's exercise of such a mechanism—the unhampered, properly-functioning exercise of which constitutes Jones's exercise of guidance control—is sufficient for Jones to be morally responsible for shooting Smith, then Jones *is* morally responsible for shooting Smith, even if determinism is true and Jones lacks ultimate responsibility for shooting Smith.[39] The compatibilist could then elaborate the sort of guidance control mechanism that Jones exercises that gives rise to Jones's "true" moral responsibility for his voluntary actions.[40] Indeed, if the Consequence Argument is sound, then it seems that a compatibilist *must* deny that: (i) "true" moral responsibility requires ultimate responsibility and (ii) "true" moral responsibility requires regulative control, not just guidance control. In this way, compatibilists avoid the foreclosure of retributivism by determinism and could, consistently, endorse both determinism and retributivism.[41]

Settling whether true moral responsibility requires ultimate responsibility, regulative control, or both exceeds this article's scope.[42] This section has argued that, *on that assumption*, determinism forecloses retributivism. In the next section (focusing on indeterminism), I will *assume* that D's ultimate responsibility is required for D to deserve punishment for committing C to consider the implications of that assumption if indeterminism is true.

V. INDETERMINISM, CONTROL, AND AGENT CAUSATION

A. Control and the "Disappearing Agent" Objection

On the assumption that D's moral responsibility requires D's ultimate responsibility, the Consequence Argument provides *prima facie* reasons to think that D's moral responsibility for blameworthy action is incompatible with determinism. But what about indeterminism? For PMR to threaten retributivism, there must also be *prima facie* reasons to think that D's moral responsibility is incompatible with indeterminism.[43] For the remainder of this section, I will assume, for the sake of argument, that indeterminism is true.

Why *might* one think that D's having ultimate responsibility for committing C at t

39. Presumably, for the mechanism's exercise to confer moral responsibility upon Jones, the mechanism also must be responsive to practical reasons available to Jones.

40. Fischer and Ravizza develop this tack. *See* JOHN MARTIN FISCHER & MARK RAVIZZA, RESPONSIBILITY AND CONTROL: A THEORY OF MORAL RESPONSIBILITY (1998).

41. As previously explained, a compatibilist could be an indeterminist. But the main point here is that even a soft determinist could be a retributivist if moral responsibility does not require ultimate responsibility.

42. I am inclined to think that ultimate responsibility *is* required for D's true moral responsibility but that regulative control *is not* required. It is because I think that ultimate responsibility is required that I am an incompatibilist. But, plausibly defending incompatibilism by defending the necessity of ultimate responsibility exceeds what I can do in this article. This article, therefore, offers no argument against compatibilists who solve PMR by denying that ultimate responsibility is required for true moral responsibility.

43. Part of my formulation of PMR is that moral responsibility for blameworthy action seems to be incompatible with indeterminism.

through the exercise of control is incompatible with indeterminism? *If*, in addition to indeterminism, the "Luck Principle" is true, *then* D's ultimate responsibility appears to be ruled out. According to the Luck Principle: "If an action is undetermined at a time t, then it's happening rather than not happening at t would be a matter of chance or luck, and so it could not be a free and responsible action."[44] D's moral responsibility and retributive desert would be precluded if *nothing*, including anything about D, determines, or controls, D's wrongdoings:

> [W]here [actions] proceed not from some cause in the characters and disposition of the person, who perform'd them, they infix not themselves upon him, and can neither redound to his honour, if good, nor infamy, if evil. . . . [T]he person is not responsible for [the actions] . . . [A]s [they] proceeded from nothing in him . . . tis impossible he can, upon [their] account, become the object of punishment or vengeance.[45]

To solve PMR, the incompatibilist could deny the Luck Principle and *coherently explain how*, when D commits crime C at t, D *could*[46] exercise control in committing C, such that:

> D is ultimately responsible for committing C at t,
> D's committing C is undetermined at t, and
> D's committing C at t is not just a matter of chance, but instead is a matter of D's exercise of the necessary sort of control, ability, or power.

A challenge to the possibility of such an explanation could be derived from Derk Pereboom's "Disappearing Agent Objection."[47] Imagine that D is deciding, a little before and up to time t, whether to commit crime C.D is motivated, a little before and up to t, by moral reasons, not to commit C, but D is also motivated, a little before and up to t, by narrowly self-interested reasons, to commit C. D's motivations are intentional, D-involving states or events that causally affect how D behaves at t. Further, imagine that

---

44. Robert Kane, *Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism*, 96 J. PHIL. 217, 217 (1999) (emphases removed).

45. DAVID HUME, A TREATISE OF HUMAN NATURE, bk. II, pt. III, § II. Robert Kane elaborates this idea: "Suppose two agents had exactly the same pasts . . . up to the point where they were faced with a choice between distorting the truth for selfish gain or telling the truth at great personal cost. One agent lies and the other tells the truth. . . . [Such] undetermined choices or actions would be 'arbitrary,' 'capricious,' 'random,' 'uncontrolled,' 'irrational,' 'inexplicable,' or 'matters of luck or chance,' and hence not free and responsible actions." Kane, *supra* note 44, at 222–23. In a similar spirit: "[i]f 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 be morally responsible for our actions. ROBERT NOZICK, PHILOSOPHICAL EXPLANATIONS 299 (1981).

46. To solve PMR, the incompatibilist must show only that incompatibilism is *consistent with* ultimate responsibility. In other words, the incompatibilist must argue only that there are possible worlds in which indeterminism is true and D is ultimately responsible. Coherently explaining such a possibility establishes consistency. Solving PMR this way, however, does not establish a libertarian position, according to which there actually are retributive facts about morally responsible agents. In other words, even if the incompatibilist solution to PMR works, all forms of retributivism might, in fact, be false.

47. *See, e.g.*, DERK PEREBOOM, FREE WILL, AGENCY, AND MEANING IN LIFE 32 (2014). Pereboom himself thinks that the incompatibilist explanation to be considered shortly, based on agent-causalism, *is* coherent but is *not likely* to be true: "Positing . . . agent-causes . . . involves no internal incoherence. There is no internal incoherence in the idea of an agent having a . . . causal power to cause her actions deliberately in such a way that her causation of her actions is not itself produced by processes beyond her control. It is unclear, however, whether we have any reason to believe that such entities exist." Derk Pereboom, *Determinism al Dente*, 29 NOÛS 21, 28 (1995).

these D-involving states or events *do not causally necessitate* D's decision. As explained previously, according to determinism, the laws of nature are deterministic.[48] According to indeterminism, however, some laws of nature are not deterministic. Some are probabilistic.[49]

Consistent with the working indeterministic assumption of probabilistic natural laws, imagine that the D-involving states or events make p the probability that D commits C at t and (1 – p) the probability that D does not commit C at t.[50] Next, imagine that D commits C at t. Although the D-involving states or events occurring before t give D's committing C at t a probability of p, *nothing* makes it the case that D commits C at t, as opposed to not committing C at t. Within the scope of the probabilities, what happens at t is *random*. It is neither D nor anything about D that brings about D's committing C at t. The agent, D, has "disappeared." D's "disappearance" at t threatens to preclude any incompatibilist explanation of how D's committing C at t is not just a matter of *chance* (bounded by probabilities), but instead is a matter of D's exercise of the necessary sort of *control*. How might the incompatibilist meet this challenge?

B. Agent Causalism

The incompatibilist might embrace *agent causalism* to meet the challenge. To clarify agent causalism, consider Christopher Franklin's "It Ain't Me Argument":

(1) An agent s self-determines a decision d only if (i) s adjudicates between his various motivations for or against d, and (ii) on the basis of this adjudicating process s determines or causes d.
(2) If the members of some set of states and events play the causal roles of (i) and (ii), then s plays the causal roles of (i) and (ii) only if s is identical to (some members of) this set of states and events.
(3) An agent is not identical to any state or event or any set of states and events.
(4) Therefore, if the members of some set of states and events play the causal roles of (i) and (ii), then s does not self-determine d.
(5) Therefore, if s self-determines d, then s, and not merely states and events, causes d.[51]

The incompatibilist could respond to the challenge posed by the Disappearing Agent Objection by endorsing step three and accepting agent causalism, according to which the agent, D, is a *substance* having a set of choice-enabling properties in virtue of which D has a certain power, which "is not characterized by any function from circumstances to effects."[52] The choice-enabling properties "make possible the direct, purposive bringing

48. *See* infra n. 26.

49. "[A] law is probabilistic if it affirms that, on the average, a stated fraction of cases displaying a given condition will display a certain other condition as well." *Law of nature*, ENCYCLOPEDIA BRITANNICA, https://www.britannica.com/topic/law-of-nature#ref285134 (last modified Mar. 20, 2014).

50. The intended interpretation of probability here is non-subjective.

51. Christopher Evan Franklin, *If Anyone Should Be an Agent-Causalist, then Everyone Should Be an Agent-Causalist*, 125 MIND 1101, 1120–21 (2016).

52. Timothy O'Connor, *Why Agent Causation?*, 24 PHIL. TOPICS 143, 145 (1996).

about of an effect by [D]."[53] "Substance" here expresses "the concept of object, or thing when this is contrasted with properties or events."[54] Substances *have* properties and are *involved in* events. Imagine that D eats a carrot named "Harvey." Harvey is a substance.[55] Orangeness and edibility are two properties that Harvey has. D's ingesting Harvey is an event that involves two substances—Harvey and D. D is not a property or an event. D has the property of having a mouth, and so on.

According to agent causalism's foil, event causalism, all causes and effects are states or events, and hence, the only kind of basic causal relation is one whose relata are states or events.[56] In contrast, according to agent causalism, although all effects are states or events, some causes are substances. There is an irreducible type of causal relation, in which D, a substance, causes a state or event. According to agent causalism,

> causation by [D] is causation by such a substance. Since a substance is not the kind of thing that can itself be an effect (though various events involving it can be) . . . [D] is in a strict and literal sense an originator of [D's] [blameworthy] decisions, an uncaused cause of them.[57]

The idea of this second, irreducible type of causal relation—substance-causation—is old.[58] According to Roderick Chisholm, Aristotle alludes to the two different types of causal relation when Aristotle states: "[a] staff moves a stone, and is moved by a hand, which is moved by a man."[59] Agent causalism depicts D as able to act in a kind of autonomous, self-determining way. Such causal autonomy requires that D have an "originating" causal control over D's actions. Having this sort of control, in turn, requires that D be a substance that can agent-cause certain events. Thus, according to agent causalism, the third step of the It Ain't Me Argument—"[a]n agent is not identical to any state or event or any set of states and events"—is true.

What would asserting that D *is identical to* a (set of) state(s) or a (set of) events amount to? We should keep retributive theories of punishment in mind when answering this question. When D is morally responsible for committing C, it is D, if anyone, who

---

53. *Id.* (emphasis removed).

54. Howard Robinson, *Substance*, STAN. ENCYCLOPEDIA OF PHIL. (last modified Feb. 3, 2014), https://plato.stanford.edu/entries/substance/.

55. In asserting that Harvey is a substance, I am bracketing potential metaphysical issues pertaining to mereological nihilism, according to which "there are no composite objects—i.e. objects with proper material parts." Gabrielle Contessa, *One's a Crowd: Mereological Nihilism without Ordinary-Object Eliminativism*, 55 ANALYTIC PHIL. 199, 199 (2014). I am ignoring any possible complications involved, for example, in holding that there are no carrots but only "atoms" arranged carrot-wise.

56. "[T]he event-causalist [contends] that the causation of events intrinsic to . . . [blameworthy] actions by [D] . . . is just a matter of 'ordinary' event-causation." John Bishop, *Agent-causation*, 92 MIND 61, 63 (1983).

57. Randolph Clarke & Justin Capes, *Incompatibilist (Nondeterministic) Theories of Free Will*, STAN. ENCYCLOPEDIA OF PHIL. (rev. ed. 2013), http://plato.stanford.edu/entries/incompatibilism-theories/ (emphasis added). The original version of this statement refers to "free," not "blameworthy," decisions.

58. There might be a *single* type of irreducible causal relation such that some relata are of a different metaphysical category (substances) than others (events). Maybe, whether an event or a substance does the causing when a state or event is brought about, the "bringing-about" itself is, irreducibly, the same.

59. Roderick M. Chisholm, *Freedom and Action*, *in* FREEDOM AND DETERMINISM 11–44 (Keith Lehrer, ed. 1966) (quoting ARISTOTLE, *Physics*, 256a). I remain neutral as to whether Aristotle intended this statement to illustrate the two types of causation under discussion. The point is simply that the idea of substance causation is an old one, possibly ancient.

deserves punishment for committing C.[60] What is the relationship between D and the wrong-making features of D's action that grounds a desert-based justification for punishing D, not someone else, for performing *that* action? Asserting that D *is identical to* a set of states or events implies a "Bundle Theory" of personal agents, such as D.

According to the Bundle Theory, D *just is* a complex event comprising experiences, intentional events, and so forth. During any time-segment or moment of D's existence, D does not "entirely exist"—only the part of D that is currently "happening" does: "there are long series of different mental states and events. . . . Each series is unified by various kinds of causal relation. . . . [A] Bundle Theorist denies the existence of persons. . . . If . . . persons are . . . separately existing things, distinct from . . . various kinds of mental states and events."[61]

So, the It Ain't Me Argument's premise that D is *not identical to* any state or event or any set of states and events entails that the Bundle Theory is false. And, it seems that the incompatibilist *must* deny the Bundle Theory to prevent the Disappearing Agent Objection from precluding any explanation of how D exercises control in committing C at t, such that: (I) D is ultimately responsible for committing C at t, (II) D's committing C is undetermined at t, and (III) D's committing C is not just a matter of chance, but instead is a matter of D's exercise of the necessary sort of control, ability, or power.

Recall that the Disappearing Agent Objection considers a scenario in which D's motivations are intentional, D-involving states or events that causally affect, without necessitating, D's committing C at t. On the Bundle Theory, D's motivations are *D-involving* intentional events in virtue of being *proper spatiotemporal parts of D*—the big event that D *is* includes them. The Bundle Theory elaborates the relationship between D and the wrong-making features of C that grounds a desert-based justification for punishing D for C as follows: C has wrong-making features. C is caused, in part, by D's motivations. D's motivations are D's motivations in virtue of being spatiotemporal parts of D, and no other agent. D is, therefore, *uniquely* related to the wrong-making features of C because it is intentional events that D, *and no other agent*, spatiotemporally comprises that "partially" cause, without necessitating, C at t.

Now, rehearse the Disappearing Agent Objection once more: Consistent with the assumption of indeterminism, imagine that the D-involving events (that D *just is*, or *comprises*) make p the probability that D commits C at t and that they make (1 – p) the probability that D does not commit C at t.[62] Next, imagine that D commits C at t. Although the series of D-involving events (that D just is, or comprises) occurring before t give D's committing C at t a probability of p, *nothing else*—indeed, *nothing at all*—makes it the case that D commits C at t, as opposed to not committing C at t. Within the scope of the probabilities, what happens at t is *random*. It is neither D nor anything about D that brings about D's committing C at t. The agent, D, has "disappeared." If D *just is* or spatiotemporally *comprises* a set of D-involving intentional events governed by probabilistic natural laws—and nothing more—then all that D or any spatiotemporal part

60. To simplify, I am ignoring the possibility of desert-based complicity liability here.

61. DEREK PARFIT, *Divided Minds and the Nature of Persons*, *in* MINDWAVES: THOUGHTS ON INTELLIGENCE, IDENTITY AND CONSCIOUSNESS 19–26 (1987).

62. The intended interpretation of probability here is non-subjective.

of D *could* do is "control" the probabilities. Within the scope of those probabilities, what happens would be *random*. The Bundle Theory invites the Luck Principle problems back into the picture.

Looking at D's control and the Bundle Theory of agents from another angle, if D *just is* a very complex event *consisting of* a set or series of causally necessitated and/or probabilistically randomized mental events—some of which form whatever chains of practical reasoning result, ultimately, in all D's (allegedly) voluntary choices and actions—then it is hard to see how D could satisfy the voluntariness condition for moral responsibility, given the assumption that moral responsibility requires ultimate responsibility. D could not "control" D's choices and actions to become ultimately responsible for them any more than a hurricane could "control" how it behaves to become ultimately responsible for its behaviour.

Thus, an incompatibilist aiming to solve PMR should accept that D is *not identical to* a series of states or events. To be the sort of thing that could exercise the necessary sort of control, ability, or power at t such that C's commission at t does not occur *randomly* (within probabilities), D must be a substance that can cause C at t without being determined to do so. D must be a substance capable of exercising *agent-causal* control at t. It is by exercising agent-causal power in bringing about C at t that D becomes morally responsible for committing C.[63]

If agent causalism is coherent, and therefore possible, the incompatibilist can solve PMR by offering a strong *prima facie* reason why D's ultimate responsibility for blameworthy action is *compatible* with indeterminism.[64] The incompatibilist can argue that retributive facts about D are possible, even if D's moral responsibility requires D's ultimate responsibility.[65] So, PMR is not an *unresolvable* puzzle about the existence of impossible things, even if ultimate responsibility is necessary for moral responsibility. There are possible worlds in which: (I) D is ultimately responsible for committing C at t, (II) D's committing C is undetermined at t, and (III) D's committing C is not just a matter of chance, but instead is a matter of D's exercise of *agent-causal* control.

To round out the discussion of the incompatibilist's "agent-causal" solution to PMR, it is worth examining the relationship between agent-causal control and the distinction between guidance control and regulative control. Reconsider Frankfurt's counterexample to the Principle of Alternate Possibilities under the assumption that Jones's moral responsibility for shooting Smith requires Jones's ultimate responsibility for doing so. Agent causalism enables an interpretation of Frankfurt's counterexample according to which: (A) Jones lacks regulative control over shooting Smith, and (B) Jones is ultimately

---

63. Some incompatibilists do not find agent causalism appealing. According to Robert Kane, incompatibilists who believe in free will sometimes:

> posit "extra factors" in the form of unusual species of agency or causation (such as noumenal selves, immaterial egos, or nonoccurrent agent causes) to account for what would otherwise be arbitrary, uncontrolled, inexplicable, or mere luck or chance. . . . Such appeals introduce additional problems of their own without . . . directly confronting the deep problems about indeterminism, chance, and luck.

Kane, *supra* note 44, at 223.

64. The possibility here is metaphysical.

65. The possibility here is metaphysical.

responsible for shooting Smith in virtue of exercising agent-causal guidance control.

The agent causalist could accept (A) because Jones is unable to do otherwise than shoot Smith, even if indeterminism is true. The explanation of *why* Jones lacks regulative control differs from the Consequence Argument's deterministic explanation, however. According to the conditions specified in Frankfurt's counterexample, the following counterfactual about Black is true: If Jones showed any signs that he would decide not to shoot Smith, then Black would intervene to override that decision by making Jones decide to shoot Smith. On the agent-causal, indeterministic interpretation, this true counterfactual is about how Black would exercise *his* agent-causal power to "override" Jones's agent-causal power should doing so suit Black's aims. Because of Black, Jones cannot access or actualise any possible world in which he does not shoot Smith. As a matter of *contingent* fact, Black would prevent Jones's access if Jones tried to actualise such a possible world.

The agent causalist could also accept (B). On the agent-causal, indeterministic interpretation, Jones is ultimately responsible for shooting Smith in virtue of exercising a form of guidance control. But, the agent causalist would deny that Jones's exercise of such guidance control can be *identical to* the operation of a properly-functioning, reasons-responsive, psychological mechanism. In addition, for exercising guidance control to confer ultimate responsibility upon Jones, his shooting Smith must "hav[e] a causal history in which [Jones] is the source of [Jones's] action in a specific way"—Jones must agent-cause the shooting.[66] By agent-causing the shooting, Jones originates it. By originating it, Jones himself is the ultimate determining cause of his shooting Smith. Jones himself is the shooting's determining cause "all the way back." By exercising agent-causal guidance control, Jones is ultimately responsible for shooting Smith, even though Jones lacks regulative control over shooting Smith.

VI. CONCLUSION

Retributive facts must exist for any form of retributivism to be true. In order for there to be retributive facts about D, D must be morally responsible for committing some crime, C. If D's moral responsibility requires D's ultimate responsibility, then D must exercise control in committing C in virtue of which D is ultimately responsible for committing C. The Consequence Argument shows that D's exercise of such control is inconsistent with determinism. And, even if determinism is false, Section V has argued that D must exercise agent-causal guidance control in committing C for D to be ultimately responsible for committing C. Therefore—If moral responsibility does not require ultimate responsibility, then both soft determinism and event-causal indeterminism are consistent with retributivism. But, on the assumption that ultimate responsibility is required for moral responsibility: (1) if determinism is true, then no retributive theory of punishment is true. And, (2) if agent causalism is false, then no retributive theory of punishment is true, even if determinism is false.

66. *See* PEREBOOM, *supra* note 47.