

The Cage We Need

Determinism, Sentencing, and the Hidden Function of Retributivism

The standard argument for reforming criminal sentencing in light of determinism runs as follows: if offenders are products of causes they did not choose, punishment for its own sake is cruelty, and sentencing should therefore become rehabilitative and forward-looking. This argument is correct about desert and wrong about consequences. It correctly identifies that retributivism's philosophical foundations — suffering owed to an offender in a way that tracks choice rather than causation — cannot survive a serious acceptance of determinism. What it misses is what retributivism has been doing institutionally that has nothing to do with desert. When a legislature genuinely accepts determinism as operative premise, it does not merely abandon an indefensible theory of punishment. It removes the only structural feature of criminal sentencing that limits what the state may do to a person to what that person has already done — rather than to what the state predicts they will do next. The humanitarian reformers have been pulling at a thread without checking what is attached to it.

The humanitarian logic is internally consistent and deserves its full force. If criminal behaviour traces causally back through neurology, developmental environment, and conditions no offender chose or authored, then desert in the philosophically robust sense is never present. Punishment inflicted as a response to choice, when there was no unencumbered choice to respond to, is not justice. It is pain. The forward-looking alternative follows: sentences should aim at rehabilitation, deterrence, and public protection, calibrated to actual risk and capacity for change. The humanitarian case earns its credibility. It also, as we will see, borrows more than it acknowledges.

P.F. Strawson's compatibilism offers the philosophical response: our practices of holding responsible are grounded not in libertarian metaphysics but in reactive attitudes — resentment, indignation, moral address — constitutive of the participant stance we adopt toward persons as agents. Determinism as abstract thesis leaves those attitudes intact. A person can believe

everything is caused and still resent an insult; the reactive attitudes are not hostage to the metaphysics.

The problem is that this argument operates at the level of the individual. It shows that a person can accept determinism while inhabiting the participant stance in ordinary life. It does not show, and was not designed to show, that a legislature building sentencing policy on the premise that offenders are products of causes to be understood and managed can maintain the participant stance institutionally. A legislature does not resent Marcus Webb. It categorises him, predicts him, and designs interventions for him. The moment legislative policy adopts the premise that criminal behaviour is a causal phenomenon, it has adopted Strawson's objective stance toward the class of persons it sentences. What follows from the objective stance is not retributivism in any form. It is case management.

Case management, followed honestly, produces outcomes retributivism prohibits. Proportionality — the principle that punishment fits the crime — is a backward-looking constraint: severity tracks the gravity of the act, and the act is bounded and finished. In a framework governed by the objective stance, what determines the appropriate intervention is not what was done but what will be done and how tractable the causal history is. The easily rehabilitated serve less; the treatment-resistant serve more — not because they deserve more, but because the framework's logic requires more intervention to change the chain. Sentence length becomes a function of modifiability. Proportionality survives, if at all, as a heuristic to be overridden when the calculus demands it.

The history of indeterminate sentencing in the United States between 1930 and 1970 is evidence about what forward-looking sentencing frameworks actually produce when implemented. Justified in explicitly rehabilitative, quasi-determinist terms — hold offenders until the causal chain is altered — the system generated sentences of extraordinary and arbitrary length. Persons convicted of minor offences served years while parole boards assessed their modifiability. This was not a failure of execution. It was the framework's logic, applied. And the logic extends further than indeterminate sentencing took it: a legislature that genuinely believed offenders are products of causes making future offending probable would have no principled basis for releasing them regardless of what they had done. Indefinite preventive detention for

predicted future offending is not a dystopian extrapolation. It is what the objective stance, consistently applied, licenses.

Before considering the most sophisticated objection, it is worth establishing precisely what the objective stance costs. Antony Duff's communicative theory of punishment identifies what is at stake: punishment as moral address to a rational agent, calling the offender to account and demanding genuine acknowledgment of wrong, requires treating that person as someone to be reasoned with rather than managed around. This communicative structure is not a refinement of punishment. It is what makes punishment — as opposed to quarantine or conditioning — a morally serious institution at all. And it requires the participant stance throughout: the sentence says something to the offender, as an agent capable of receiving and responding to moral address. Strip away the participant stance, and nothing communicative remains. You do not send a moral message to a system of causes. You intervene in it.

This matters because the objection to the consequentialist drift of determinist sentencing typically invokes human dignity as the constraint that prevents the worst outcomes. But dignity, in the relevant sense, is not a free-standing value that can be bolted onto any framework. It is what the participant stance looks like when formalised — the institutional expression of treating persons as agents whose authorship of their acts commands a specific form of response. Duff's point is that you cannot retain the substance of that constraint while abandoning the stance that generates it. A dignity constraint in an objective-stance framework is formal rather than foundational: it can be asserted, but it cannot be derived from what the framework takes itself to be doing.

Hart separates the general justifying aim of punishment — which is consequentialist — from the principles governing its distribution among offenders, which are deontological. On this account a determinist legislature can adopt rehabilitation and public protection as its aims while retaining proportionality as an independent distributive constraint, doing genuine normative work that the consequentialist aim cannot override.

The vulnerability is specific, not the general instability that afflicts any hybrid framework. A Rawlsian hybrid grounds its deontological constraints in the separateness of

persons — a principle with its own independent foundation, derived from a prior account of persons as self-authoring agents that does not depend on the reactive attitudes for its authority. That constraint can survive the abandonment of desert because it was never grounded in desert to begin with.

Hart's proportionality principle has no such independence. Its entire normative content derives from the idea that what the offender did, as their act, is the morally relevant fact that fixes what may now be done to them. That idea is a participant-stance response: resentment and moral address expressed as law. When Hart strips desert from his justifying aim, he does not merely remove one element of retributivism while retaining another. He removes the only source of authority his distributive constraints possessed. They float free — present in the framework, derivable from nothing within it. Whether other hybrid frameworks face the same problem depends on whether their foundational constraints genuinely escape the participant stance or merely defer to it — a further argument this essay cannot make within its scope, but one the Rawlsian case suggests is harder to win than it appears.

This is not a comfortable conclusion, and it would be dishonest to present it as one. It requires defending the structural function of a framework whose philosophical foundations are genuinely indefensible — and it offers no clean resolution, because the claim here is precisely that those foundations and that function cannot be separated. This is not a pragmatic argument for retributivism, a claim that it produces good outcomes and should be preserved instrumentally. It is a claim about architecture: retributivism constrains the state because it answers to the past act, and it answers to the past act because it takes seriously the offender's authorship of that act. Abandon the authorship, and the constraint goes with it. There is no version of the framework that keeps the cage and discards the philosophy.

A legislature that accepted determinism as operative premise would not merely be adopting a more honest theory of punishment. It would be dismantling the only structure that kept its claim on persons tethered to what they had done rather than open-ended toward what they might do. The sentence ends when the debt is paid, because the debt is defined by the act and the act is over. There is no debt in a consequentialist framework. Only risk. And risk does

not expire. The cage of retributivism is philosophically embarrassing. It is also the only cage with a door.