DRAFT COPY

DO NOT DISTRIBUTE

BIAS AT THE SURFACE OR THE CORE? A COMMENT ON THE PSYCHOLOGY OF THE TRIAL JUDGE

Eyal Aharoni

Is judicial decision making steeped in cognitive bias? And which way does it lean in the perennial rivalry between emotion and reason? In *The Psychology of the Trial Judge*, Morris Hoffman draws on a studious reading of the scientific literature and on his real-world experience as a trial judge to examine these difficult questions. In this commentary, I attempt to raise some additional questions that stem from his analysis, as applied to the legal justification for criminal punishment.

By way of background, the question of judicial bias represents a longstanding debate between two formidable opponents: The rationalist perspective argues that reason plays the leading role in moral judgment (Turiel, 1983; May, 2018) whereas decision bias is a supporting actor at best. The implication of this perspective is that education and expertise should tend to reduce decision bias. The intuitionist account makes the opposite case—that emotion leads, while reason plays little to no causal role. Like an emotional dog with a rational tail, it would be a mistake to think the tail is the primary driver of the dog's trajectory (Haidt, 2001). If human decision making is fundamentally emotional and intuitive, decision bias, even among legal experts, would tend to be far-reaching and relatively immutable.

So, which is it? Ultimately, Hoffman's analysis resists this dichotomy, favoring instead a more nuanced narrative between these extremes. That narrative acknowledges that even expert judges may not be immune to various cognitive biases, but expertise does seem to afford at least some protection. So let's not throw the baby out with the bathwater, says Hoffman, because beneath the veneer of psychological biases lies a basically rational core.

To evaluate this conclusion, it might help to deconstruct the question into two forms. The weak form asks whether judges are in some ways affected by bias. The answer here is a decisive yes. There is ample empirical evidence of this, which Hoffman characterizes fairly. Scholars simply disagree about how much, and that question will be directly informed by further research, as Hoffman rightly acknowledges.

The strong form asks whether judicial decision making, being an expression of human decision making more generally, is fundamentally rational or emotional and intuitive. This is a much harder question to

answer, and it is where Hoffman's interpretation might be slightly more optimistic than my own. One reason for my reluctance about claims of judicial rationality lies in the observation that we, as a society, do not have a clear conceptual understanding of why we punish criminals, and this lack of understanding makes it effectively impossible to punish rationally. There is, of course, a rich history of scholarship on the legal justifications for punishment, namely deontological theories of retribution and consequentialist theories of behavior control, with retribution recognized as the dominant justification in current U.S. law (American Law Institute § 1.02(2), 2017). But this framework is deeply flawed, and these flaws pose additional challenges for the prospect of a rational law.

To defend this proposition, let us review what rationality means in this context. To engage in *rational* decision making implies, at the very least, some strategic attempt to optimize a behavior with respect to some conceivable goal state. For example, if it is important to me that my child gets good grades, then hiring a tutor might count as a rational strategy in the pursuit of my goal. (Certainly, there are other definitions of rationality, such as that of logical reasoning or the pursuit of self-interest, but these forms of rationality exhibit even higher standards than means-goal rationality.) At minimum, being goal-directed necessitates a prospective relationship between the means and the end. But this is exactly opposite of what deontological theories of retribution prescribe. Instead, these theories assert that violators deserve punishment for what they did, regardless of its bearing on future behavior (Packer, 1968). So, being explicitly retrospective in nature, what observable goal state is retribution designed to optimize for?

To obviate my point, there is an illustrative case of a Denver teen, Darrell Havens, who was paralyzed from the neck down after being shot by a police officer during the theft of a car. He would never be able to steal a car again, and yet he was sentenced to 20 years in prison. (The total taxpayer burden was estimated to the tune of $4 million, had his sentence not been cut short by an apparent suicide (Prendergast, 2017)). And for what—to set an example that "if you're thinking about using paralysis as a way to get away with crime, you better think twice?" Surely, there are better examples by which to send a deterrent message to society.

When people do bad things but are no longer dangerous, and when their example does not lend itself to a clear deterrent message, most people still demand punishment—but we have a hard time articulating why (Aharoni & Fridlund, 2011). Our retributive impulses do not seem ruffled by the fact that no amount of punishment can actually change the past. Instead, theories of retribution offer rhetoric about expressing condemnation for its own sake, or as a means to restoring a moral balance (Kant, 1785/1998; Duff, 2001), but they do not clearly define what a moral balance is let alone how or when it has been restored (see Greene, 2014; Hart, 2008). This deontological hand waving confirms to me that retribution,

as acknowledged by some of its own proponents, is not goal-directed at all. And if it is not goal-directed, it is fundamentally incompatible with the enterprise of rationality. The implication is that participants in criminal justice institutions are carrying out the bidding of enigmatic, emotionally charged motivations that make insistent demands for the delivery of harm to others yet offer no clear goal definition or way to optimize for some goal. So, I worry that when we punish on retributive grounds, we are using the offender as a scapegoat to help us manage our own evasive and intractable moral emotions. I worry that retributive sentiments are expressions of psychological bias.

If the desire for retribution itself is a bias, then bias does not just grow at the periphery of our criminal punishment system—it is at the core. Granted, there are other (i.e., consequentialist) justifications to lean on. But neither legal theory nor practice offer ways for us to reconcile conflicts between the two types of justifications. They seem fundamentally incompatible, and the law lacks a higher order theory prescribing when one should override the other. So, how can we expect judges to punish rationally, if we cannot even provide them with a coherent answer to the basic question: what is the primary goal of the punishment?

This is where it becomes instructive to consider, from a scientific perspective, how human decision making developed into its current form. Research in the evolutionary sciences has begun to uncover quite specific answers to the question of what goals our retributive motivations are optimized to serve. After examining these descriptive reasons, we as a society can more cogently evaluate whether we endorse or condemn them.

To make sense of retributive behavior, we need to consider the environment in which our modern minds adapted. In that environment, our ancestors didn't have state police forces and prison facilities, so they needed efficient, interpersonal solutions for managing threats and leveraging social relationships for their own benefit. Indeed, people who were motivated (by retributive emotions) to punish transgressors could have derived a variety of fitness advantages, whether by deterring or directly incapacitating one's oppressor, or deterring opportunistic onlookers, for example. Through these mechanisms, the expression of a strong commitment to punish transgressors can advertise the punisher's social capital, discourage future threats, and facilitate cooperation, as Hoffman and others have contended (Aharoni & Fridlund, 2013; Delton & Krasnow, 2017; Fehr & Gächter, 2002; Fehr & Fischbacher, 2004; Fiddick, Cosmides, & Tooby, 2000; Frank, 1988; Hoffman, 2014; Krasnow et al., 2012; Krasnow et al., 2016; Pedersen, McAuliffe, & McCullough, 2018; Petersen et al., 2010; Price, Cosmides, & Tooby, 2002; Trivers, 1971).

Owing to these advances in evolutionarily-informed scholarship, we can, for the first time, articulate clear, observable (albeit descriptive) goals to be achieved by retributive punishment. Notice that these goals have a striking resemblance to the standard consequentialist justifications for punishment like deterrence

and incapacitation, except for a key difference: emotion. According to evolutionary theories of punishment, punitive emotions like moral outrage (which are irrelevant by traditional consequentialist accounts) evolved because they provided a computationally inexpensive way of motivating punishment behavior. In ancestral environments, as with retaliatory actions observed in other animals (e.g., Clutton-Brock & Parker, 1995), you shouldn't have to deliberate your way to a punishment decision; that would be far too slow and costly. Retributive emotions solved that problem using automation. From an evolutionary perspective, this feature makes retributive motives begin to sound an awful lot like a strategic bias.

If retributive punishment decisions are fundamentally biased, is that necessarily a bad thing? If it evolved to reduce threats and increase cooperation, isn't that just the kind of bias we need in our justice system? Yes and no. Certainly, we can take inspiration from the scientific research on punishment, and ask whether we want to optimize punishment for evolved goals. Some of these (like incapacitation and deterrence) seem like legitimate societal goals to me, while others (like jockeying for social status) do not. But therein lies a major caveat in the extent to which we ought to lean on our retributive biases as guides to punishment: they evolved to benefit *individuals*, not societies. Indeed, research on our ancestors' darker behaviors have obviated the conclusion that traits that have conferred reproductive advantages to individuals—such as adultery, child maltreatment, and homicide—are a poor guide for a functional and moral *society* (Daly & Wilson, 1988). So, if we embrace retribution in criminal punishment, I submit that we carry a burden to prove that it is not doing more harm than good to society.

Ultimately, debates about the justification for punishment have much to learn from ongoing research and scholarship. Traditional notions of these justifications (namely the deontological versus consequentialist theories) are at a stalemate. At first glance, scientific perspectives on the evolution of our motivations for punishment might seem to complicate that relationship even more. Since discoveries about our ancestral heritage are purely descriptive, they cannot, by themselves, justify how we ought to live.

Yet, in addition to providing an explanation for why human beings seek retribution, research on our evolved psychology of punishment can be of use in another important way: it provides a new framework for unifying rival legal theories. This framework suggests that retributive and consequentialist motives for punishment are not orthogonal as legal philosophy would suggest—they are just different levels of analysis for describing a single psychological phenomenon. When a moral violation evokes strong feelings of moral outrage, these feelings evolved precisely because they motivated our ancestors to act on them, and such actions conferred important benefits upon punishers, such as protection from future harm. So, ironically, retribution is a proximate representation of an evolved psychology whose ultimate function

was consequentialist (Aharoni & Hoffman, in press; Cushman, 2015). Once endowed with the impulse to punish, you need not be conscious of the advantageous function of the impulse in order to benefit from it. You just need to submit to the impulse. Retributive theorists seem to be in the business of providing post hoc justifications for this impulse whereas consequentialist theorists require than any justification for such an impulse must explicitly fulfill a legitimate prospective goal. So when retributive and consequentialist theorists talk past each other, they may be simply engaging in different levels of analysis of the same phenomenon. Understanding the relationship between these levels can obviate the differences between the goals targeted by our intuitive impulse to punish and those targeted by our modern social values.

Scholars have argued that we can leverage this newfound scientific knowledge to make clearer, more goal-directed, collective decisions (Jones, 2000). For example, if we were to collectively decide that the dominant goal of punishment should be to protect public safety, then we don't need to argue about whether it is moral to assign lengthy prison terms to quadriplegic car thieves. Instead, we could conclude that this action is not an efficient means to our collective goal. Indeed, it could siphon away scarce funds from the punishment of more dangerous individuals. But if we never question the legitimacy of our retributive impulses, we have no way to prevent them from getting the better of us.

Asking whether judges are biased has both a short answer and a long one. The short answer, on which Hoffman and I closely agree, is: sure, kind of. The long answer—about which thoughtful minds can disagree—taps a much deeper question about the present and proper roles of emotion and reason in our justice system. Evolutionary scholarship casts doubt on the integrity of the folk psychology on which the law is based, but it also offers the law a powerful tool for reconciling the historically incompatible justifications for punishment by showing that what feels like a retributive impulse in the proximate sense might have design features for fulfilling particular fitness goals for our ancestors. Our foremost goal as a society, in my view, should be to decide, with the aid of rigorous research and scholarship, when we are better off fulfilling those evolved goals or protecting ourselves from them.

References

Aharoni, E. & Fridlund, A.J. (2011). Punishment without reason: Isolating retribution in lay punishment of criminal offenders. *Psychology, Public Policy, and the Law 18*(4), pp. 599-625.

Aharoni, E. & Fridlund, A.J. (2013). Moralistic punishment as a crude social insurance plan. In T. Nadelhoffer (Ed.), *The Future of Punishment.* New York: Oxford University Press, pp. 213-229.

Aharoni, E. & Hoffman, M.B. (in press). Evolutionary psychology, jurisprudence, and sentencing. In T. Shackelford (Ed.), *The SAGE Handbook of Evolutionary Psychology.* London: SAGE Publications.

American Law Institute (2017). *Model penal code.* Philadelphia, Pa. The American Law Institute. Retrieved on October 22, 2019 from https://archive.org/stream/ModelPenalCode_ALI/MPC

Chase, W.G. & Simon, H.A. (1973). Perception in chess. *Cognitive psychology 4*(1), pp. 55-81.

Clutton-Brock, T.H. & Parker, G.A. (1995). Punishment in animal societies. *Nature 373*(6511), p. 209.

Cushman, F. (2015). Punishment in humans: From intuitions to institutions. *Philosophy Compass 10*(2), pp. 117-133.

Daly, M. & Wilson, M. (1988). *Homicide*. New Brunswick, NJ: Transaction Publishers.

Daly, M. & Wilson, M. (2005). Carpe diem: Adaptation and devaluing the future. *The Quarterly Review of Biology 80*(1), pp. 55-60.

Delton, A.W., & Krasnow, M.M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior 38*(6), pp. 734-743.

Duff, A. (2001). *Punishment, Communication, and Community.* New York: Oxford University Press.

Fehr, E. & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior 25*(2), pp. 63-87.

Fehr, E. & Gächter, S. (2002). Altruistic punishment in humans. *Nature 415*(6868), p. 137.

Fiddick, L., Cosmides, L. & Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition 77*(1), pp. 1-79.

Frank, R.H. (1988). *Passions Within Reason: the strategic role of the emotions*. New York: WW Norton & Co.

Greene, J.D. (2014). *Moral tribes: Emotion, reason, and the gap between us and them.* New York: Penguin.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review 108*(4), p. 814.

Hart, H.L.A. (2008). *Punishment and responsibility: Essays in the philosophy of law.* Oxford: Oxford University Press.

Hoffman, M.B. (2014). *The Punisher's brain: The evolution of judge and jury.* New York: Cambridge University Press.

Jones, O.D. (2000). Time-shifted rationality and the Law of Law's Leverage: Behavioral economics meets behavioral biology. *Nw. UL Rev. 95*, p. 1141.

Kant, I. (1998). Kant's groundwork of the metaphysics of morals. In P.

Guyer's (Ed.) *Critical Essays on the Classics*. Lantham, MD: Rowman and Littlefield. (Original work published in 1785)

Krasnow, M.M., Cosmides, L., Pedersen, E.J. & Tooby, J. (2012). What are punishment and reputation for?. *PLOS ONE 7*(9), e45662.

Krasnow, M.M., Delton, A.W., Cosmides, L. & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science 27*(3), pp. 405-418.

May, J. (2018). *Regard for reason in the moral mind.* Oxford: Oxford University Press.

Packer, H. (1968). *The limits of the criminal sanction*. Stanford, Cal.: Stanford University Press.

Pedersen, E.J., McAuliffe, W.H. & McCullough, M.E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General 147*(4), p. 514.

Petersen, M.B., Sell, A., Tooby, J. & Cosmides, L. (2010). Evolutionary Psychology and Criminal Justice: A Recalibrational Theory of Punishment and Reconciliation. In H. Høgh-Oleson (Ed.), *Human Morality & Sociality: Evolutionary & Comparative Perspectives.* New York: Palgrave MacMillan.

Prendergast, A. (2017, April 27). *The Strange Death of Darrell Havens, Prisoner Who Battled the System.* Retrieved on October 12, 2019 from https://www.westword.com/news/darrell-havens-paralyzed-prisoner-who-battled-the-system-has-died-9006082

Price, M.E., Cosmides, L. & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior 23*(3), pp. 203-231.

Trivers, R.L. (1971). The evolution of reciprocal altruism. *The Quarterly review of biology 46*(1), pp. 35-57.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge, England: Cambridge University Press.