RETRIBUTION AND THE EVOLUTION OF HUMAN PUNISHMENT

Joseph Brammall

A thesis submitted in partial fulfilment of the requirements for the degree of Bachelor of Arts (Honours) in Philosophy

Supervised by Professor Peter Godfrey-Smith

University of Sydney

October 2017

Acknowledgments

First, I would like to thank my supervisor, Professor Peter Godfrey-Smith, for guiding me patiently through the philosophy of biology, the ins-and-outs of human evolution, and beyond. Your diligent, naturalistic, and truth-seeking approach to philosophy was inspiring, and I feel privileged to have been your supervisee. Thank you for all of the discussions about fascinating philosophical topics, and for your attentive and invaluable advice throughout the year. Thank you also to Yarran Hominh, Anthony Hooper, Jake Dye, and Oliver Seville, for your extremely helpful correspondence early on.

My flat-mates, Tim and Elly, thank you for putting up with my stress and lack of presence during intense periods. Max and Nick, thanks for listening to my philosophical rants, and for always looking out for me. Rupert—my fellow honours victim—thanks for the daily trips to subway, the over-caffeinated freak-outs, the insane conversations about obscure academic topics, and for your companionship and constant help with all things honours-related. Thanks also to Andy, Ania, Allie, Claud, Grace, Kate, Phil, Trav, and Sarah, for putting up with me reading about evolutionary biology while in Japan, for allowing me to drop off the Snapchat map while studying, and for being a continuous cheer squad.

Thanks to my sister, Lily, for our endless trips to and from uni, and for your constant stream of humour and kindness—it kept me sane. Finally, I would like to thank my Mum. I would not have been able to complete this thesis without her. I can't even begin to explain all the different ways you helped me over the last year—but thank you for being an endless source of strength and support.

Contents

[T]he conviction persists—though history shows it to be a hallucination—that all the questions that the human mind has asked are questions that can be answered in terms of the alternatives that the questions themselves present. But in fact, intellectual progress usually occurs through sheer abandonment of questions together with both of the alternatives they assume—an abandonment that results from their decreasing vitality and a change of urgent interest. We do not solve them: we get over them. Old questions are solved by disappearing, evaporating, while new questions corresponding to the changed attitude of endeavor and preference take their place. Doubtless the greatest dissolvent in contemporary thought of old questions, the greatest precipitant of new methods, new intentions, new problems, is the one effected by the scientific revolution that found its climax in the "Origin of Species."

—John Dewey, 1910.[1]

[1] "The Influence of Darwin on Philosophy." in *The Influence of Darwin on Philosophy and Other Essays*. New York: Henry Holt and Company (1910): 1–19.

1. Introduction

Social scientists have long considered punishment to be one of the most promising candidates for a 'human universal'—a pan-cultural behaviour common across all human societies: from tribes of hunter-gatherers, to flourishing industrial civilisations.[2] Due its ubiquity, punishment has, in the last decade or so, become a major topic of interest for evolutionary theorists looking to explain the origins of modern human behaviour in terms of natural selection.[3] For centuries prior to this, however, philosophers have been questioning the rationale behind punishment.

The main philosophical question about punishment is justificatory. Is it morally permissible to punish? And if so, why? There are two main kinds of philosophical answers to this question.[4] Consequentialists maintain that punishment is justified because it brings about good consequences.[5] And retributivists maintain that punishment is justified because wrongdoers inherently deserve to be punished.[6] Recent psychological evidence suggests that people endorse both retributivism and consequentialism in principle. But in practice, experiments have shown people are motivated to punish solely for retributive concerns, and by strong emotional reactions to wrongdoing.[7]

This paper has two primary aims. The first is to provide an evolutionary explanation for why punishment is retributively motivated. And the second is to re-evaluate the philosophical theory of retributivism, from an evolutionary perspective. The method of the paper will be naturalistic, in the sense that I will rely heavily on scientific research, and will avoid positing

---

[2] Morris Hoffman and Timothy Goldsmith, "The biological roots of punishment." *Ohio State Journal of Criminal Law*. 1 (2003): 627–641, at 627.
[3] Samuel Bowles and Herbert Gintis, *A Cooperative Species: Human Reciprocity and its Evolution.* (Princeton University Press, 2013), 4.
[4] David Wood, "Punishment: consequentialism." *Philosophy Compass* 5, no. 6 (2010): 455–469, at 456.
[5] Ibid.
[6] Ibid.
[7] Kevin Carlsmith and John Darley, "Psychological aspects of retributive justice." *Advances in Experimental Social Psychology* 40 (2008): 193–236, at 209.

anything empirically unprovable. I will draw from a variety of academic disciplines, including anthropology, economics, evolutionary biology, psychology, and philosophy—but I will present my argument in an analytic philosophical style.

In chapter 2, I define punishment, lay some conceptual groundwork, examine a body of empirical evidence about punishment. In chapter 3, I investigate what it means to explain a behaviour in terms of evolution by natural selection, and I construct an explanatory framework with which to present an evolutionary exposition of human punishment. Following this, I present an evolutionary exposition of punishment in humans. I argue that the biological function of punishment in ancestral human groups was to sustain cooperation. I contend that an emotionally-driven, retributive psychology evolved as the principal mechanism for punishment, because it was an effective and reliable instrument for motivating humans to behave in ways that increased group-level reproductive success. Finally, I explain why modern humans are motivated to punish for emotionally-driven retributive reasons. My explanation, in essence, is that the psychological mechanism for punishment that evolved in ancestral humans persists in the psychology of modern humans.

In chapter 4, I turn to the second aim of this paper: a re-evaluation of retributivism. First, I expound Greene's theory about the psychological processes that underpin moral judgement, and link his theory to punitive-decision making.[8] Next, I introduce and defend Greene's empirical explanation for the existence of deontological philosophy. Combining my evolutionary exposition of human punishment with Greene's account of deontological moral philosophy, I argue that retributivism is not, first and foremost, a philosophical theory of punishment. Instead, it is a post hoc rationalisation of the emotionally driven retributive psychology that evolved in ancestral humans for the function of motivating punishment. I conclude that retributivism should no longer

[8] Joshua Greene, "The secret joke of Kant's soul." *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* 3 (2008): 35–79.

be considered a justificatory theory of punishment, and that it should instead be considered a linguistic expression of an evolutionarily ancient psychological predisposition to punish.

2. What is Punishment?

Philosophical interest in punishment usually stops at legal punishment. Legal punishment is state-administered, and is delivered by penal institutions such as the criminal justice system. Plainly, however, not all punishment is legal. Parents punish misbehaving children, and teachers punish disruptive students. Across both legal and non-legal contexts, punishment can be defined as the intentional imposition of some sort of cost, hardship or burden in response to a believed wrongful act or omission, as an expression of condemnation or censure of that wrongful act or omission.[9] For scientific purposes, however, this definition comprises an uneasy mixture of behavioural and psychological criteria. Nakao and Machery define punishment in more general terms, as "any action that harms another organism, where that action is elicited by some specific harmful action (or trait) performed by the punished organism."[10] From a biological point of view, this definition has the advantage of applying to both humans and non-human species. But it has the disadvantage of including actions that do not intuitively count as punishment, such as self-defence.

Building on Nakao and Machery's approach, Cushman defines punishment as: "actions that harm another organism for the purposes of modifying their behaviour."[11] Cushman characterises punishment not as a class of behaviour, but instead as a functionally-defined natural kind of behaviour. I am partial to Cushman's definition, but it is worth recognising that it counts

---

[9] Alec Walen, "Retributive Justice." *The Stanford Encyclopedia of Philosophy* (Winter 2016), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/justice-retributive/>.

[10] Hisashi Nakao and Edouard Machery, "The evolution of punishment." *Biology and Philosophy* 27, no. 6 (2012): 833–850, at 834–5.

[11] Fiery Cushman, "Punishment in humans: From intuitions to institutions." *Philosophy Compass* 10, no. 2 (2015): 117–133, at 123.

quite a wide range of human actions as punishment. These actions can be grouped into three forms of punishment: direct, exclusion-based, and reputation-based.[12] Direct punishments are actions that cause direct harm or a reduction in the payoffs for an individual, such as hitting, hurting, seisure of property, or fining. Exclusion-based punishments are actions that cause harm by removing an individual from some beneficial interaction or group, such as banishment, ostracism, or shunning. And reputation-based punishments are actions that cause harm by expressing attitudes of condemnation or disapprobation; either publically, as in denouncement, or privately, as in gossiping.

As well as identifying the types of human behaviour that count as punishment, it is important to note that two kinds of relationship can occur between wrongdoers, victims of wrongdoing, and punishers.[13] The first kind of relationship is second-party punishment, which occurs when an original victim of wrongdoing delivers punishment directly to a wrongdoer, as in the case of Alice punching Bob for stealing her food. And the second kind of relationship is third-party punishment, which occurs when punishment is not delivered to a wrongdoer by the original victim of wrongdoing, but is instead delivered by a third party on the victim's behalf, as in the case of Carol punching Bob, for stealing Alice's food.

2.1. The Philosophy of Punishment

The main philosophical question about punishment is justificatory.[14] Why is it morally permissible to punish, if at all? Answers to this question are diverse, but a useful way to think about the question is to distinguish between two super-categories of justificatory theories: consequentialist theories, and retributivist theories.[15]

---

[12] Chandra Sripada, "Punishment and the strategic structure of moral systems." *Biology and philosophy* 20, no. 4 (2005): 767–789, at 779.
[13] Cushman, "Punishment," 124.
[14] Wood, "Punishment: consequentialism," 456.
[15] Morris Hoffman, *The Punisher's Brain: The Evolution of Judge and Jury*. (Cambridge University Press, 2014), 334.

Consequentialist theories hold that punishment is justified because of its future benefits.[16] In other words, consequentialists are only interested the pragmatic rationale behind punishment.[17] One of the early formulations of the consequentialist justification of punishment was offered by Jeremy Bentham.[18] For Bentham, punishment was only justifiable insofar as it contributed to the greater good of society. And the way punishment achieved this, according to Bentham, was by deterring criminals from committing crime in the future. This function of punishment is now referred to as 'special deterrence', and it is usually contrasted with 'general deterrence', the idea that punishment should function to discourage not only criminals from committing crime, but also deter law-abiding members of society.[19] Other influential consequentialist justifications of punishment include 'incapacitation', the idea that criminal incarceration protects law-abiding citizens from danger by curtailing the freedom of criminals to commit crime;[20] 'rehabilitation', the idea that punishment should function to safely re-integrate criminals into society; and 'restitution', the idea that punishment in some sense eliminates the gains made by a criminal, thereby restoring the material and psychological wellbeing of the victim of the crime, the victim's family, or both.[21]

Retributivist theories, by contrast, hold that punishment is justified *irrespective* of its future benefits. For retributivists, the rationale of punishment is that guilty wrongdoers intrinsically deserve to be punished.[22] Contemporary formulations of retributivism are diverse, but they share three common elements. The first is the idea that punishment is an intrinsically deserved response to wrongdoing, which is often known is the principle of 'desert'.[23] The second is a sense of fairness, or justice, in determining the harshness of punishment. Wrongdoers should only be punished in

---

[16] Wood, "Punishment," 456.
[17] Greene, "Secret joke," 50.
[18] Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation (Chapters I–V).* (Blackwell Publishing Ltd, [1789] 1972).
[19] Wood, "Punishment," 456.
[20] Wood, "Punishment," 462.
[21] Hoffman, *Punisher's Brain*, 344.
[22] David Wood, "Punishment: nonconsequentialism." *Philosophy Compass* 5, no. 6 (2010): 470–482, at 470.
[23] Ibid.

proportion with their blameworthiness, and the degree of harm they cause. This idea is often known as the principle of 'proportionality'.[24] And the third element is the idea that punishment is only appropriate when wrongdoers are believed with a reasonable degree of confidence to be guilty.[25] Early philosophical formulations of retributivism are traceable to Hegel[26] and Kant,[27] but prototypes of retributivism pre-date contemporary Western philosophy by millennia. For example, retributive ideas are found in the Old Testament, as in the Mosaic law 'an eye for an eye.'[28]

Although consequentialist and deontological theories of punishment are often held to be in tension,[29] it is important to recognise that not all justificatory theories of punishment fall squarely into either the retributivist or consequentialist category. Pluralistic theories of punishment, such as those offered by H. L. A. Hart[30] and John Rawls,[31] combine both consequentialist and retributive elements. Generally, however, the distinction between retributivist and consequentialist theories is a useful way to think about the justificatory debate, because it captures two very different ways of thinking about the value of punishment. Consequentialists appeal to the instrumental value of punishment,[32] whereas retributivists appeal to its intrinsic value.[33]

---

[24] Ibid., 471.
[25] Hugo Bedau and Erin Kelly, "Punishment," *The Stanford Encyclopedia of Philosophy* (Fall 2015), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2015/entries/punishment/>.
[26] Georg W. F. Hegel, *Elements of the Philosophy of Right*. Trans. N. B. Nisbet, Ed. Allen Wood. (Cambridge: Cambridge University Press, [1821] 1991). 119–131.
[27] Immanuel Kant. *The Metaphysics of Morals*, M. Gregor (trans.), (New York: Cambridge University Press [1797] 1991), 141–142.
[28] Exodus 21: 23–25, *The Holy Bible: King James Version*. Texas: National Publishing Company (2000).
[29] Hoffman, *Punisher's Brain*, 344.
[30] H. L. A. Hart, "Prolegomenon to the Principles of Punishment." *Proceedings of the Aristotelian Society* 60 (1959–1960): 1–26.
[31] John Rawls, "Two Concepts of Rules." *Philosophical Review* 64 (1955): 3–32.
[32] Wood, "Punishment: consequentialism," 457.
[33] Wood, "Punishment: nonconsequentialism," 470.

2.2. The Psychology of Punishment

In the last two decades, scientific evidence about the psychological motives of punishment has emerged from research on punitive decision-making. In general, the aim of this work was to determine whether people's attitudes about punishment aligned with the retributivist rationale for punishment, or whether people were instead motivated to punish for consequentialist reasons.[34]

An important initial finding was that when people were asked in an abstract way about the rationale of punishment, they mentioned both retributivist *and* consequentialist reasons for punishment.[35] In other words, people gave answers that appealed to the future benefits of punishment (such as deterrence and incapacitation), and they also gave answers that reflected the idea that proportional punishment is a deserved response to culpable wrongdoers.[36] A natural hypothesis to draw from this finding, is that people are motivated to punish for a combination of consequentialist and retributivist reasons. As it turned out, however, people's verbal reports of their motives for punishment were a poor reflection of their actual punitive behaviour.

In one study, subjects were instructed to make punitive judgements in simulated criminal cases, and were given the option of learning information pertaining to desert and proportionality (the magnitude of harm, the perpetrator's intent, and extenuating circumstances), or learning information pertaining to the future benefits of delivering punishment (frequency of the crime, likelihood of detection, and publicity of the crime).[37] The first set of information was designed to represent retributivist considerations, and the second was designed to represent consequentialist considerations. It was found that subjects readily asked for information that was relevant from a

---

[34] Carlsmith and Darley, "Psychological aspects," 209.
[35] Bernard Weiner, Sandra Graham, and Christine Reyna, "An attributional examination of retributive versus utilitarian philosophies of punishment." *Social Justice Research* 10, no. 4 (1997): 431–452, at 450.
[36] Kevin Carlsmith, John Darley, and Paul Robinson, "Why do we punish? Deterrence and just deserts as motives for punishment." *Journal of personality and social psychology* 83, no. 2 (2002): 284–299, at 295.
[37] Kevin Carlsmith, "The roles of retribution and utility in determining punishment." *Journal of Experimental Social Psychology* 42, no. 4 (2006): 437–451, at 448.

retributivist standpoint, and showed very little interest in information that was relevant from consequentialist standpoint. Moreover, retributively-relevant information gave people more confidence in their punitive decisions, while consequentially-relevant information had the opposite effect.

In a different study, it was found that subjects' punitive judgements were highly sensitive to retributively-relevant factors, such as the severity of the offence, but that consequentially-relevant factors, such as the likelihood of re-offence, had a minimal influence on their judgments.[38] To test this effect, some subjects were explicitly instructed to make their punitive-decisions in a consequentialist way, by focusing on deterrence related factors. Even then, however, those subjects were unable to punish in a genuinely consequentialist way. They decided the severity of punishment according to retributivist factors, while randomly increasing the severity of punishment for the sake of deterrence in a way that did not accurately track factors that were relevant to deterrence.

Another set of studies focused on the origin of the psychological motivation to punish. A key finding was that punitive decision-making was closely associated with negative emotions on the part of the punisher, such as anger, and disapproval.[39] Carlsmith, Darley and Robinson coined the term 'moral outrage' to describe the sentiment that people feel in reaction to wrongdoing.[40] They found that the extent to which people reported feelings of moral outrage correlated not only with the blameworthiness of the wrongdoer and the degree of harm caused, but that moral outrage was also good predictor of the severity with which people decided to punish.[41] The idea that punitive decision-making is closely related to emotion was also found in brain-imaging studies. In one study, it was found that the severity of punishment in response to violations of trust was

---

[38] John Darley, Kevin Carlsmith, and Paul Robinson, "Incapacitation and just deserts as motives for punishment." *Law and Human behavior* 24, no. 6 (2000): 659–683.
[39] Daniel Kahneman, David Schkade, and Cass Sunstein, "Shared outrage and erratic awards: The psychology of punitive damages." *Journal of Risk and Uncertainty* 16, no. 1 (1998): 49–86.
[40] Carlsmith, Darley, and Robinson, "Why do we punish," 284.
[41] Ibid.

correlated with the level of activity in the caudate nucleus, a brain region associated with emotion, motivation and reward.[42] And a different study found that emotionally graphic descriptions of harmful acts boosted brain activity in the amygdala, a brain region associated with emotional reaction and decision-making, and that amplified punishment severity correlated with increased amygdala activity.[43]

2.3. The Central Question

Empirical evidence shows that people endorse both retributivist *and* consequentialist reasons for punishment in the abstract, but pay next to no attention to consequentialist considerations when punishing. Instead, people were motivated to punish for retributive reasons, which depended on factors such as the severity of wrongdoing, and the blameworthiness of the wrongdoer. And punitive decision-making was closely connected to negative emotional reactions, such as moral outrage. In short, people do not have a good understanding of why they punish at all.

It is now widely-accepted among cognitive scientists, philosophers, and psychologists, that human reasoning is dictated to a significant extent by automatic processes, biases, and unconscious heuristics. Daniel Kahneman, a Nobel-prize winning psychologist, describes human reason as comprising two distinct processing systems: one that is fast, instinctive and emotional, and another that is slower, more deliberative, and more logical.[44] And Jonathan Haidt's well-known work on 'moral dumbfounding' in the psychology of morality showed that moral judgements are largely derived from quick, instinctive intuitions that people cannot rationally explain (for instance, the intuition that incestuous intercourse is repugnant), rather than careful reasoning about right and

[42] Dominique De Quervain, et al., "The neural basis of altruistic punishment." *Science* 305, no. 5688 (2004): 1254–1258.
[43] Michael Treadway, et al., "Corticolimbic gating of emotion-driven punishment." *Nature Neuroscience* 17, no. 9 (2014): 1270–1275.
[44] Daniel Kahneman, *Thinking, fast and slow*. (Macmillan, 2011), 20–23.

wrong.[45] Psychologists are increasingly moving towards the view that punitive decision-making is also motivated by fast, intuitive psychological processes, rather than slow, rational processes.[46] Carlsmith, for instance, argues that the moral outrage and retributive motives that dominate punitive decision-making are the conscious registration of intuitive emotional reactions towards wrongdoing.[47] And Treadway et al. claim that people rely heavily on emotional heuristics when making punitive decisions.[48]

This empirical story about punishment raises important questions. Consequentialist reasons for punishment (such as deterrence, incapacitation, and rehabilitation), play a clear and undeniable role in the justification of punishment. The deterrence of criminal behaviour, the incapacitation of dangerous offenders, and the reintegration of prisoners into society are fundamental bases for institutional incarceration. And in non-institutional contexts, the logic of punishment is plainly connected to beneficial future consequences. When an adult punishes a misbehaving child, the aim is to discourage the child from behaving similarly again in the future.

But if the future benefits of punishment factor so significantly into its rationale, then why is it the case that consequentialist considerations are completely absent from the psychological processes that occur when people deliver punishments? When asked about the justification of punishment in the abstract, people think about consequentialist and retributivist reasons. But why do people suddenly forget their consequentialist considerations when asked to make a real punitive decision? In other words, why is punitive decision-making psychologically motivated by emotionally-driven retributivist intuitions, and not by considerations about the beneficial things that punishment brings about?

---

[45] Jonathan Haidt, "The emotional dog and its rational tail: a social intuitionist approach to moral judgment." *Psychological review* 108, no. 4 (2001): 814–834.
[46] Kevin Carlsmith et al., "Psychological aspects," 211.
[47] Ibid., 212.
[48] Treadway et al., "Corticolimbic gating," 1270.

In recent decades, an answer to this question has emerged from a combination of disciplines, including evolutionary biology, sociobiology, game-theoretic economics, anthropology, and evolutionary psychology. In the next chapter, I will present an interdisciplinary explanation for why human punishment is motivated by retributive emotion, and not by consequentialist considerations.

3. Why is Punishment Retributively Motivated?

In the same way that early 20th century Western philosophy underwent a 'linguistic turn', characterised by a tendency to examine the relationship between philosophy and language[49]—it could reasonably be said that the social sciences are in the middle of an 'evolutionary turn.' Disciplines as diverse as anthropology, economics, jurisprudence, linguistics, and psychology, are increasingly explaining the evolution of human behaviours and mental processes in terms of natural selection.[50] (To clarify, I mean natural selection in the strict, biological sense, where *homo sapiens* is regarded as a species that evolved from natural origins; rather than natural selection in the fallacious social-Darwinian sense, where ethics are derived from nature). A wide range of human traits have been subject to evolutionary analysis, but one human trait that has received a great deal of explanatory attention is 'altruistic cooperation'.[51] Humans have a behavioural predisposition to help other humans, including non-kin, and even despite the risk of personal disadvantage.[52] How and why did humans come to be this way? Recent work in evolutionary theory suggests that punishment plays a central role in the answer.[53]

---

[49] Ronald Allen and Brian Leiter, "Naturalized epistemology and the law of evidence." *Virginia Law Review* (2001): 1491–1550, at 1493–4.
[50] Ian Gough, et al., "Darwinian evolutionary theory and the social sciences." *Twenty-First Century Society* 3, no. 1 (2008): 65-86, at 66.
[51] Bowles and Gintis, *A Cooperative Species*, 4.
[52] Ibid.
[53] Ibid., 148–150.

In this chapter, I will examine what it means to explain the evolution of a trait in terms of natural selection. Following this, I will develop an explanatory framework with which to present a biological exposition of the evolution of human punishment. The aim of this exposition is to provide a comprehensive answer to the question that I raised in §2.3. Namely, why is punitive decision-making psychologically motivated by emotionally driven retributivist intuitions, and not by considerations about the beneficial things that punishment brings about?

3.1. Evolution by Natural Selection

To start with, a brief outline of evolution by natural selection will be of use. All populations of organisms continually undergo change with respect to their behaviour, bodies, and physiology (biologists call these changes 'variations').[54] Variations occur randomly, and are caused at a microscopic level by genetic mutations, and by the recombination of parents' chromosomes during sex.[55] Sometimes variations will bestow an organism with a reproductively advantageous trait, allowing that organism to produce more offspring than its conspecifics (biologists call this 'differential reproduction').[56] If the advantageous trait is genetically heritable, then in many cases it will be passed on to the offspring of the organism.[57] As this pattern repeats itself, the advantageous trait can proliferate, and cause evolutionary change in entire populations of organisms.[58]

There are ongoing debates about many evolutionary concepts, (such as 'fitness' and 'adaptation'),[59] but for present purposes, these debates can be overlooked. It is important to note, however, that evolutionary change is not driven by any higher purpose or end-goal.[60] Instead, it is

[54] Peter Godfrey-Smith, *Philosophy of biology*. (Princeton University Press, 2013), 30.
[55] Ibid., 39.
[56] Ibid., 30.
[57] Ibid.
[58] Ibid., 42.
[59] Ibid., 33.
[60] Ibid., 60.

driven by incremental reproductive advantages that accumulate over the lifecycles of individual organisms.

3.2. Functions in Biology

Some species of horned-lizards (*Phrynosoma*) exhibit an unusual behaviour known as 'autohaemorrhaging', or 'reflex bleeding'.[61] When a horned-lizard is threatened by a predator, it will increase the blood pressure in small sinus chambers under its eye sockets, and squirt a jet-like stream of blood from one of its eyes towards the predator, up to distances of 1.2 meters. Predators are quickly frightened off by this bizarre exhibition. The function of this autohaemorrhaging behaviour seems fairly evident: it protects the lizard from predation. But *why* is this the function of autohaemorrhaging? In everyday parlance, the term 'function' is often used in a sense roughly synonymous with the terms 'purpose' or 'role.' In biology, however, the concept of function has a more technical meaning, which is associated with evolutionary theory.[62] Seeing as the biological concept of function will feature regularly throughout this chapter, it is important to examine the concept in detail. A good place to start is with Wright's general theory of functions,[63] as it elucidates the way functions operate in biology. Wright's theory, expressed formulaically, is as follows:

> "The function of X is that particular consequence of its being where it is which explains why it is there."[64]

The first point to take from Wright's theory is that the concept of function in biology is associated with a particular kind of explanation. Specifically, biological functions explain 'why X is there.' Wright's terminology is somewhat obscure, but in the context of biological functions, 'X'

---

[61] Wade Sherbrooke and George Middendorf, "Blood-squirting variability in horned lizards (Phrynosoma)." *Copeia* 2001, no. 4 (2001): 1114–1122, at 1115.
[62] Peter Godfrey-Smith, "A modern history theory of functions." *Noûs* 28, no. 3 (1994): 344–362.
[63] Larry Wright, "Functions." *The philosophical review* 82, no. 2 (1973): 139-168.
[64] Larry Wright, *Teleological explanations: An etiological analysis of goals and functions*. (University of California Press, 1976), 78.

stands for any functionally-characterised biological entity, including behaviours such as autohaemorrhaging. And the expression 'is there' refers to the *existence* of the functionally-characterised entity.[65] The key to explaining why any functionally-characterised biological entity exists, is to recall how evolution by natural selection operates. Heritable traits that increase reproductive success are passed on and proliferate in populations.[66] The process of evolution by natural selection explains why functionally-characterised biological entities exist, and this means that the existence of any functionally-characterised biological entity in an organism's phenotype can only be explained by looking backwards in time, to the evolutionary history of that entity.[67]

Applied to the horned-lizard case, we find that the proper functional explanation for autohaemorrhaging is that autohaemorrhaging enabled horned-lizards to avoid being eaten, to survive, and experience more mating opportunities. This reproductive advantage caused the trait for autohaemorrhaging to be passed on and proliferated in horned-lizard populations—and this explains why autohaemorrhaging exists in the horned-lizard phenotype.

The second point to take from Wright's theory is that functional explanations in biology only concern the aspect of a biological entity that increases an organism's reproductive success.[68] The previous example will help to illustrate the point. In addition to the deterrence of predators, several ancillary effects are caused when a horned-lizard autohaemorrhages. The lizard's blood volume decreases, it is temporarily blinded by blood, and its eye socket swells. The reason it would be strange to explain the function of autohaemorrhaging in terms of any of these ancillary effects, is that none explain why autohaemorrhaging exists in the horned-lizard phenotype. The functionally-relevant aspect of autohaemorrhaging is the capacity to squirt blood at predators, because the ability to avoid predation allowed horned-lizards to survive, to experience more mating opportunities, and to pass on the trait for autohaemorrhaging. In other words, the function of a

---

[65] Godfrey-Smith, "Modern history," 353.
[66] Godfrey-Smith, "Philosophy," 30.
[67] Godfrey-Smith, "Modern history," 351.
[68] Wright, *Teleological explanations*, 78.

biological trait is not just any effect caused by the trait. Rather, the function of a biological trait is the aspect of the trait that explains its existence in the organism's phenotype.

Wright's theory is a good representation of the way functions operate in biology, but it does not account for the fact that a trait may be selected for more than one function over the course of its evolutionary history.[69] To account for this, evolutionary biologists distinguish between 'adaptations' and 'exaptations'.[70] Adaptations are traits that were shaped by natural selection for their current function. Exaptations, by contrast, are traits that acquire functions for which they were not originally selected.[71] In the flightless ancestors of modern birds, such the *Archaeopteryx*, feathers were probably selected for insulation.[72] In modern birds, however, feathers were selected for the ability to assist with flight. For example, think of the long, stiff feathers of a peregrine falcon (*Falco peregrinus*).[73] It would be strange to consider insulation the main function of a peregrine falcon's feathers, because the function of a trait does not necessarily depend on the original reason it was selected. Instead, the function of a trait is the reason for a trait's recent maintenance in a selective context.[74] A useful way to flesh out this idea is to distinguish between the 'ancient' and 'modern' evolutionary history of a biological trait.[75] The ancient evolutionary history of a trait involves the original selective forces that built the structure of the trait, which account for its initial emergence, and its original function.[76] The modern history of a trait, by contrast, involves the reason a trait was selected for in its most recent selective episode.[77] Functional explanations in biology concern only the 'modern' evolutionary history of traits.

---

[69] Godfrey-Smith, "Modern history," 357.
[70] Ibid.
[71] Ibid.
[72] Ibid.
[73] Ibid.
[74] Ibid., 358.
[75] Ibid., 357.
[76] Ibid., 356.
[77] Ibid.

Taking these clarifications into consideration, Wright's general theory of functions can be reformulated as follows to represent the way functions operate in biology:

> "The function of a biological entity is the aspect of that entity which explains why the entity currently exists."

3.3. The Proximate–Ultimate Distinction

In 1961, Ernst Mayr introduced a biological distinction between 'proximate' causes—the immediate physiological responses of an organism and their organs to factors in the environment; and 'ultimate' causes—the events or factors in the evolutionary history of an organism that shaped the trait via natural selection over many generations.[78] The proximate–ultimate distinction is frequently used as way of understanding evolutionary explanations,[79] but there has been substantial debate over: (a) whether the proximate–ultimate distinction is appropriate for theoretical use,[80] and (b) the sense in which evolutionary biologists actually use the distinction.[81] Seeing as the proximate–ultimate distinction is often to organise evolutionary explanations, it is appropriate to indicate where I stand on these issues.

One interpretation of the way biologists use the proximate–ultimate distinction is that the expression 'proximate causes' is used to refer to factors that operate within the life-span of an individual organism; whereas the expression 'ultimate causes' is used to refer to processes that occurred before the organism was conceived, and which shaped its genome via natural selection.[82] Francis affirms the importance of this conceptual distinction, but he rejects use of the proximate–ultimate distinction to label it.[83] The correct way to refer to the distinction, he argues, is in terms

---

[78] Ernst Mayr, "Cause and effect in biology." *Science* 134, no. 3489 (1961): 1501-1506, at 1503.
[79] Nick Davies, John Krebs and Stuart West, *An Introduction to Behavioural Ecology, 4th ed.,* (Wiley-Blackwell: Chichester UK, 2012), 2.
[80] Brett Calcott, "Why the proximate-ultimate distinction is misleading, and why it matters for understanding the evolution of cooperation." *Cooperation and its evolution* (2013): 249–263, at 250.
[81] Richard Francis, "Causes, proximate and ultimate." *Biology and Philosophy* 5, no. 4 (1990): 401-415, at 403.
[82] Ibid., 404.
[83] Ibid., 405.

of 'ontogenetic factors' and 'phylogenetic factors."[84] Ontogenetic factors relate to 'ontogeny'—the development of individual organisms, or with the development of the anatomical or behavioural features of organisms from the earliest stage to maturity.[85] And phylogenetic factors relate to 'phylogeny'—the evolutionary development and diversification of a species or group of organisms, or of a particular feature of an organism.[86]

A different interpretation of how biologists use of the proximate–ultimate distinction, offered by Haig, is that the expression 'proximate causes' is used to refer to explanations that appeal to the local mechanisms that cause a behaviour or trait, and the expression 'ultimate causes' is used to refer to explanations that appeal to the biological function of a behaviour or trait.[87] As was the case with Francis, Haig rejects the proximate–ultimate distinction as a label for the mechanism–function distinction, on the grounds that the term 'ultimate' is too ambiguous.[88] Francis also takes specific issue the term 'ultimate', claiming that it has connotations of exhaustiveness, fundamentality, and superiority, all of which are inappropriate for theoretic contexts.[89] In their final analyses, Francis and Haig both deem the proximate–ultimate distinction too ambiguous for theoretical use. I agree with this view, and will therefore avoid using the proximate–ultimate distinction in my biological exposition of human punishment.

3.4. Tinbergen's Four Kinds of Explanation

A different approach to understanding evolutionary explanations is to distinguish between four kinds of explanation that can be given for any organism's behaviour.[90] On this approach,

---

[84] Ibid.
[85] Davies, Krebs and West, *Introduction*, 2.
[86] Ibid.
[87] David Haig, "Proximate and ultimate causes: how come? and what for?" *Biology & Philosophy* 28, no. 5 (2013): 781–786, at 782–3.
[88] Ibid., 781.
[89] Francis, "Causes," 403.
[90] Nikolaas Tinbergen, "On aims and methods of ethology." *Zeitschrift für Tierpsychologie,* 20, (1963): 410–433, at 411.

which was introduced in 1963 by Nikolaas Tinbergen, any question of the form 'why does organism *O* exhibit behaviour *Φ*?' can be answered in terms of:

(1) The function or 'survival value' of *Φ*.[91]
(2) The evolutionary history of *Φ*.
(3) The development of *Φ* throughout the life of *O*.
(4) The physiological mechanisms and physical stimuli that cause *O* to exhibit *Φ*.[92]

Tinbergen's approach is widely-used in contemporary biology,[93] and it has a clear advantage over the proximate–ultimate distinction, in that it subsumes and differentiates between both of the conceptual distinctions identified by Haig and Francis. The distinction between phylogenetic factors and ontogenetic factors is represented in (2) and (3), and the distinction between function-based explanations and mechanism-based explanations is represented in (1) and (4).[94] There are, however, a few issues with Tinbergen's framework.

Tinbergen treats functional explanations (1) as different in kind to explanations that appeal to evolutionary history (2). Yet as I previously argued, functional explanations in biology must appeal to evolutionary history.[95] Tinbergen's demarcation between (1) and (2) is at odds with the fact that functional explanations in biology necessarily appeal to evolutionary history. Recall, however, that functional explanations in biology concern only the modern evolutionary history of a trait (i.e., the recent selective maintenance of a trait). The ancient evolution history of a trait (the original selective forces that built the structure of a trait and which account for its initial emergence) may be irrelevant to the trait's function, as in the case of the peregrine falcon's feathers.

---

[91] Tinbergen uses the terms 'survival value' and 'function' interchangeably.
[92] Godfrey-Smith, "Modern history," 351.
[93] Ibid.
[94] Pat Barclay and Mark Van Vugt, "The evolutionary psychology of human prosociality: Adaptations, byproducts, and mistakes." in *The Oxford Handbook of Prosocial Behavior* (Oxford Library of Psychology, 2015), 37-60, at 37.
[95] Godfrey-Smith, "Modern history," 351.

To account for this, it is worth making two small refinements to Tinbergen's framework. Functional explanations concern only the *current* function of a trait (i.e., the function that continues to be maintained by natural selection), and explanations in terms of evolutionary history concern only the origin and phylogenetic distribution of a trait.

Tinbergen developed his explanatory framework to clarify the scientific method of ethology: the study of the evolution of (non-human) animal behaviour in terms of natural selection. As such, he did not account for the fact that in humans, behaviours can be triggered by psychological mechanisms, such as emotional urges, or practical reasoning.[96] It is worth making one final refinement to Tinbergen's framework, to account for the fact that human behaviour is often caused by psychological mechanisms. Taking into account all of the above, I will employ the following explanatory structure in my evolutionary exposition of punishment in humans:

(1) The current selective function of punishment.
(2) The origin and phylogenetic distribution of punishment.
(3) The development of punishment throughout the lives of individual humans.
(4) The psychological mechanism(s) that cause humans to exhibit punishment.

3.5. Punishment in Humans: An Evolutionary Exposition

It is an uncontroversial fact that humans have a predisposition to punish other humans.[97] In formal evolutionary theory, the human predisposition to punish is treated as one component of a construct known as 'strong reciprocity.'[98] Strong reciprocity involves altruistic preferences, which are sentiments that place a positive value on the beneficial outcomes of one's actions for others, even when those actions bear personal costs.[99] Strong reciprocity motivates individuals to

[96] Barclay and Van Vugt, "The evolutionary psychology," 39–40.
[97] Hoffman, *Punisher's Brain*, 311.
[98] Bowles and Gintis, *A Cooperative Species*, 20.
[99] Ibid., 4.

sacrifice their own payoffs in order to cooperate with others, to reward the cooperation of others, and to sacrifice their own payoffs in order to punish non-cooperation.[100]

In modern human societies, punishment largely is delivered by institutions such as the criminal justice system.[101] Yet is not necessary to account for these institutions in what follows, because it is highly unlikely that sophisticated punitive institutions existed during the period in which the human predisposition for punishment evolved.

3.5.1. The Selective Function of Punishment

The human predisposition to punish most likely evolved in the Late Pleistocene epoch (the period from between about 126 and 12 thousand before the present), in mobile foraging bands of hunter-gatherers living in Sub-Saharan Africa.[102] In this early stage of human history, life as a hunter-gatherer was extremely dangerous. Lethal intergroup warfare was common, resulting in the frequent dispersion and eradication of human groups.[103] And a harsh, volatile climate meant that material resources were scarce.[104] Bowles and Gintis hypothesise that in these hostile conditions, groups comprising individuals with a predisposition to altruistically help their fellow group members experienced increased group-level reproductive success.[105] The idea that natural selection operates at both the individual-level *and* the group-level is known as 'multi-level selection.'[106] According to Bowles and Gintis, within-group altruistic cooperation conferred significant reproductive advantages for ancestral humans, in that it allowed humans to form hunting parties, to share food and child rearing responsibilities, to team-up in combat, and to collaboratively acquire and defend territory.[107]

---

[100] Ibid.
[101] Ibid., 92.
[102] Ibid., 2.
[103] Ibid., 8.
[104] Ibid., 97.
[105] Ibid., 51.
[106] Ibid., 53.
[107] Ibid., 196.

The function of punishment, (the aspect of punishment that explains why a predisposition to punish currently exists in the human phenotype), was to solve a specific problem that arose in cooperative groups.[108] Cooperation generated reproductive advantages in ancestral humans groups, but it also gave individuals the opportunity to exploit the benefits of cooperation, by free-riding (receiving the gains of cooperation, but contributing nothing in return) and by violating pro-social norms (breaking rules of behaviour designed to maximise cooperation).[109] According to Cushman, the function of punishment was to allow early humans to modify the non-cooperative behaviour of their fellow group members,[110] by teaching them not to exploit the benefits of within-group cooperation.[111] In other words, punishment sent a very strong message to non-cooperators. "If you cause harm to the group, or refuse to perform actions that benefit the group, then you will be harmed in return."

The behaviour-modification function of punishment has been demonstrated in a large body of evolutionary mathematical models,[112] and has repeatedly been confirmed in game-theoretic experiments, most notably a series of 'public goods games' designed by Fehr and Gachter.[113] When individuals were permitted to punish the non-cooperative behaviour of others, it was found that group-average cooperation increased significantly.[114]

3.5.2. The Origin and Phylogenetic Distribution of Punishment

The exact evolutionary origin of human punishment is difficult to pinpoint, because the predisposition to punish is thought to be an exaptation, rather than an adaptation. One theory is

---

[108] Ibid., 4.
[109] Ibid., 31.
[110] Cushman, "Punishment in humans," 123.
[111] Ibid.
[112] Robert Boyd et al., "The evolution of altruistic punishment." *Proceedings of the National Academy of Sciences* 100, no. 6 (2003): 3531-3535.
[113] Ernst Fehr and Simon Gächter. "Altruistic punishment in humans." *Nature* 415, no. 6868 (2002): 137-140.
[114] Individuals opted to altruistically invest monetary units into a 'public goods fund' as opposed to keeping possession of monetary units out of self-interest.

that the trait for punishment evolved via a two-stage exaptation process, in which a primitive pan-mammalian aggression system that evolved for the original function of resource protection was selected a second time, for the function of preservation of positions in dominance hierarchies; and third time, for the function of behaviour-modification.[115]

In terms of phylogenetic distribution, the main puzzle in explaining how punishment initially emerged and proliferated in ancestral human populations stems from the fact that punishment was highly costly for hunter-gatherers to deliver.[116] Punishers risked the possibility of physical harms, such as retaliation, or social harms, such as resentment.[117] Social costs were particularly damaging in the Late Pleistocene, as group membership was more or less essential to survival.[118] Because of this, Bowles and Gintis suggest that the benefits of punishment were characterised by increasing returns to scale, meaning that the total cost of punishing declined as the number of punishers increased.[119] They hypothesise that a specific set of conditions were necessary for the early emergence and proliferation of a predisposition to punish. First, punishment needed to be contingent on the number of group members who were willing to participate.[120] This type of collective action is known in biology as 'quorum sensing', whereby individuals do not behave as a group until there are enough of them to effectively bring about an action.[121] And second, in order for punishment to be contingent, it needed to be coordinated and organised between group members prior to delivery.[122] For coordination to be effective, individuals likely needed a strong capacity for cheater detection, and norms of truthful information sharing.[123]

---

[115] Tim Clutton-Brock and Geoffrey Parker, "Punishment in animal societies." *Nature* 373, no. 6511 (1995): 209–216, at 210.
[116] Bowles and Gintis, *A Cooperative species*, 148.
[117] Ibid., 24.
[118] Ibid., 35.
[119] Ibid., 149.
[120] Ibid., 161.
[121] Ibid., 149.
[122] Ibid., 150.
[123] Ibid., 163.

3.5.3. The Development of Punishment

All human traits develop under the influence of a combination of genetic and cultural factors.[124] Some traits are passed on purely genetically, but by and large, human development is significantly affected by learning and enculturation.[125] In stable conditions, humans can culturally transmit information and artificial environments over thousands of generations.[126] Academic knowledge, institutions, languages, skills, social norms, and technology are all examples culturally-transmitted factors.[127]

It is possible to roughly differentiate between the genetic basis of a predisposition to punish, and the culturally transmissible factors that influence the development of the punitive predisposition. The genetically heritable component of punishment is a pre-linguistic psychological impulse to harm individuals who exhibit non-cooperative behaviour. And the culturally transmitted factor is the set of social norms that individuals learn throughout their life, which will influence their beliefs about punishment, including the method by which punishment should be delivered, and which actions constitute non-cooperative behaviour. Social norms can be transmitted vertically (from parents to offspring), obliquely (from non-parental members of a parent's generation) or horizontally (between individuals of a similar age); but individuals may also construct or modify their beliefs about punishment independent of the influence of social norms.[128]

It is crucial to note, however, that over many generations of human life, genetic and cultural factors continuously interact.[129] Culturally transmitted factors create artificial

---

[124] Herbert Gintis, "Gene–culture coevolution and the nature of human sociality." *Philosophical Transactions of the Royal Society B: Biological Sciences* 366, no. 1566 (2011): 878–888, at 878.
[125] Bowles and Gintis, *A Cooperative Species*, 126.
[126] Gintis, "Gene–culture coevolution," 880.
[127] Ibid.
[128] Bowles and Gintis, *A Cooperative Species*, 15.
[129] Gintis, "Gene–culture coevolution," 878–888.

environments in which previously non-advantageous mutations become advantageous, are selected for, and proliferate.[130] In evolutionary theory, this feedback loop is known is as 'gene–culture co-evolution.'[131] Moreover, culturally transmitted traits can themselves be selected for or against, according to the reproductive success that they confer to their carriers. This process is commonly known as cultural selection.[132] In §3.2, I described the biological concept of function only in terms of a genetic selection-mechanism. And in §3.5.1, I described the selective function of punishment only in genetic terms. But it is important to acknowledge that many functionally-characterised humans traits, including punishment, are maintained not just by genetic selection, but by a *mixture* of selection-mechanisms. In what follows, when I refer to the 'selective function' or 'function' of punishment, I recognize the influence that cultural selection and gene–culture co-evolution will have on the selective maintenance of functionally-characterised human traits. Although I am confident that punishment was influenced by an interplay between genes and culture, I am uncertain as to the details of how culturally-transmitted social norms influenced the genetic basis of a psychological predisposition to punish.

3.5.4. The Psychological Mechanism for Punishment

In 1971, Robert Trivers proposed the idea that an individual-level psychological mechanism known as 'moralistic aggression' evolved in ancestral humans, for the function of protecting vulnerable altruistic individuals from exploitation by non-cooperators.[133] According to Trivers, the modus operandi of moralistic aggression was to generate a negative emotional response in reaction to non-cooperative behaviour, which would psychologically register as

---

[130] Ibid., 882.
[131] Ibid.
[132] Ibid., 879.
[133] Robert Trivers, "The evolution of reciprocal altruism." *The Quarterly review of biology* 46, no. 1 (1971): 35-57, at 49.

feelings of anger, contempt, or indignation.[134] These negative sentiments would then motivate altruistic individuals to:

> "…educate the unreciprocating individual by frightening him with immediate harm or with the future harm of no more aid; and in extreme cases, perhaps, to select directly against the unreciprocating individual by injuring, killing, or exiling him."[135]

Trivers' work on moralistic aggression was an early conceptualisation of the view that *negative emotion* is the psychological mechanism for punishment in humans. This view has since been endorsed by evolutionary theorists such as Bowles and Gintis, who note that:

> "…ethically motivated outrage, what Robert Trivers called 'moralistic aggression,' is a plausible motivation for the strong reciprocators' punishment of defectors."[136]

Empirical evidence suggests the same conclusion. In a series of public goods experiments, Fehr and Gächter hypothesised that: "free-riding may cause strong negative emotions among the cooperators and these emotions, in turn, may trigger their willingness to punish the free riders."[137] This hypothesis was confirmed, as it was found that the severity with which people punished instances of free-riding matched self-reports of the intensity of the negative emotions they felt in reaction to observations of free-riding.[138]

It worth mentioning that positive emotions (as opposed to negative emotions) may also play a causal role in human punishment. It is often observed that people find enjoyment and satisfaction in successfully punishing individuals who behave non-cooperatively.[139] I am unsure as to how positive emotions might generate punishment, but one promising view is positive emotions act as a kind of reward-mechanism, motivating individuals to punish in anticipation of the

---

[134] Ibid.
[135] Ibid.
[136] Bowles and Gintis, *A Cooperative Species*, 168.
[137] Fehr and Gächter, "Altruistic punishment," 139.
[138] Ibid.
[139] Bowles and Gintis, *A Cooperative Species*, 38.

satisfaction that follows from punishing.[140, 141] For present purposes, however, I will assume that negative emotion constitutes the principal psychological mechanism for punishment.

3.6. The Selective Function of Negative Emotion

The aim of this chapter was to explain why punitive-decision making is motivated by retributive emotion and not by consequentialist considerations. To meet this aim, a final matter needs to be addressed. Given the likelihood that punishment was both potentially costly and a conducive to reproductive success in ancestral human groups, it is somewhat surprising that humans evolved an emotion-based mechanism for punishment, rather than a reasoning-based mechanism. Mental capacities such as the ability to problem-solve, to engage in rational decision-making, and to perform complex reasoning tasks, are widely-considered to have played a major role in the evolutionary success of humans.[142] Given this fact, the question arises: why was emotion—an ostensibly irrational feature of human psychology—selected for as the mechanism for punishment?

The most compelling answer to this question is that emotion was selected for its long-term strategic role in increasing group-level reproductive success.[143] On this view, emotion act as a form of 'commitment device', motivating individual humans behave in ways that may seem irrational or unproductive in the short-term, but which increase reproductive fitness in the long-run.[144] Negative emotion such as anger and indignation reliably motivated humans to punish, [145] because these feelings were strong, visceral, and very difficult to cognitively override. In other words, negative emotions *forced* ancestral humans to punish. Negative emotions was also able to bypass reason-based decision-making processes about punishment, which may have inhibited long-term

---

[140] Greene, "Secret joke," 71.
[141] De Quervain, et al., "The neural basis," 1256.
[142] Robin Dunbar, "The social brain hypothesis." *Brain* 9, no. 10 (1998): 178-190.
[143] Cushman, "Punishment in humans," 123.
[144] Robert Frank, *Passions within reason: the strategic role of the emotions.* (WW Norton & Co, 1988), 51.
[145] Ibid., 51–53.

reproductive success.[146] As Cushman writes:

> "In order for an emotional mechanism to work, it must circumvent reasoning in order to perpetrate a local irrationality that achieves a larger strategic end."[147]

The other advantage of an emotion-based mechanism is that emotions are "fast and frugal."[148] Hypothetically, natural selection could have endowed humans with a purely reason-based mechanism for punishment, whereby all punitive decisions were made on the basis of rational cost-benefit analyses. However, a reason-based mechanism for punishment would have been slow, cognitively expensive, and may not have always achieved the most selectively advantageous outcome. A far more efficient mechanism, from an evolutionary perspective, was for individuals to simply experience an immediate surge of negative emotions in reaction to non-cooperative behaviour.

It is a somewhat disconcerting thought that the urge to punish wrongdoing was built into the psychology of humans by natural selection to secure reproductive success. For some reason, it is more natural to believe that intuitive reactions to wrongdoing are governed by reason, rather than emotion. But the fact that human punishment is emotionally-driven is typical of a more general pattern which applies to a wide range of human emotions and impulses. As Greene writes,

> "Nature doesn't leave it to our powers of reasoning to figure out that ingesting fat and protein is conducive to our survival. Rather, it makes us hungry and gives us an intuitive sense that things like meat and fruit will satisfy our hunger . . . when nature needs to get a behavioural job done, it does it with intuition and emotion wherever it can."[149]

3.7. Answering the Central Question

In §2.2, I presented a body of empirical evidence about the psychological motives of

---

[146] Bowles and Gintis, *A Cooperative Species,* 187.
[147] Cushman, "Punishment in humans," 126.
[148] Isaac Wiegman, "The Evolution of Retribution: Intuitions Undermined." *Pacific Philosophical Quarterly*, 98, (2017): 193–218.
[149] Greene, "Secret Joke," 60.

punitive decision-making. The key finding from this body of evidence was that when people were asked abstractly about the justification of punishment, they provided a mixture of consequentialist and retributivist reasons. But while actually delivering punishment, people abandoned their consequentialist ideologies, and punished in a solely retributive manner. An equally important finding was that punitive judgment was driven by (and proportionate to) moral outrage. In §2.3, I raised the question: why is punitive decision-making psychologically motivated by emotionally-driven retributivist intuitions, and not by considerations about the beneficial things that punishment brings about? In this chapter, I presented a biological story about the evolution of punishment in humans. Punishment was selected for the function of sustaining cooperation in groups of hunter-gatherers, by modifying the non-cooperative behaviour of free-riders and norm-violators. And negative emotional reactions, such as anger and indignation, were selected for as the mechanism to motivate punishment among ancestral humans.

In light of this evolutionary story, it is now possible to answer the question that I raised in §2.3. Punitive decision-making is psychologically motivated by emotionally-driven retributivist intuitions because negative emotions were selected for as the psychological mechanism for punishment in ancestral humans. Negative emotion was a selectively advantageous mechanism for punishment, because emotions reliably caused humans to punish. Automatic reactions of anger, contempt and indignation in response to wrongdoing were more effective and economical than reason-based cost-benefit analyses about the possible consequences of punishment. This explains why punitive decision-making is triggered by emotionally-driven retributive intuitions, and not by consequentialist considerations.

Despite the absence of consequentialist reasoning in punitive decision-making, it is crucial to recognise that punishment evolved precisely *because* of the reproductively advantageous consequences that it brought about—specifically, increased within-group cooperation. Retributive

emotions were instrumental to reproductive success, and this accounts for the fact that humans are, as Cushman puts it, "psychological retributivists."[150]

To be more accurate, the notion that humans have 'retributive' emotions and intuitions is somewhat misleading, because these emotions and intuitions are only 'retributive' in the sense that they seem to reflect retributivist theories of punishment. It is not difficult, however, to imagine why punitive emotions and intuitions evolved to track 'retributive' factors, such as the wrongdoer's intention, extenuating circumstances, and the amount of damage caused by the wrongful act. Disproportionate punishment was costly to wrongdoers, punishers, and the group. Overly-severe punishment could cause excessive harm to the wrongdoer, which would in turn increase the likelihood of retaliation or resentment. And overly-lenient punishment was costly to the group, insofar as it that it would fail to effectively modify non-cooperative behaviour.

4. Re-evaluating Retributivism

In the previous chapter, I presented an evolutionary exposition of punishment. My aim in doing so was to explain why human punishment is psychologically motivated by retributive emotion, rather than consequentialist reasoning. The answer I arrived at, in essence, was that retributive emotion evolved as the psychological mechanism for punishment in ancestral humans, and that this psychological mechanism persists in modern-day humans. In §2.1, I mentioned the two main kinds of theories that philosophers offer when justifying punishment: consequentialist theories, and retributivist theories. Consequentialists hold that punishment is justified because of its future benefits. And retributivists, by contrast, hold that punishment is justified because guilty wrongdoers intrinsically deserved to be punished, irrespective of any future benefits, that might be brought about. How does the preceding evolutionary story about punishment bear on philosophical theories of punishment? And, more specifically, what is the relationship between

[150] Cushman, "Punishment in humans," 124.

'retributive' emotions, which evolved as the psychological mechanism for punishment in humans, and the philosophical theory of retributivism?

In effort to answer these questions, I will examine Joshua Greene's theory of the two psychological processes that underpin moral judgement, and his evidence-based explanation for the existence of deontological moral philosophy.[151] Greene's work will serve as a point of contact between the evolution of human punishment and the philosophical theory of retributivism.

4.1. Greene's Theory of Moral Judgement

In normative theory, deontology is roughly the idea that actions are morally right or wrong just in virtue of their internal features, and that right actions have a reason-providing authority (often expressed in terms of duties, or rules), that make them morally required.[152] Consequentialism, by contrast, is roughly the idea that whether an action is morally right or wrong depends only on the consequences of that action.[153] Greene's theory about moral judgement, in a nutshell, is that deontology and consequentialism are *not* first and foremost normative theories.[154] Instead, according to Greene, deontology and consequentialism are two dissociable and pre-theoretic modes of moral thinking, which correspond with two natural kinds of psychological process. Deontological thinking, according to Greene, is generated by automatic, emotional processes; whereas consequentialist thinking is generated by deliberative, cognitive processes.[155]

Greene's theory might seem counterintuitive at first sight, because deontology is generally associated with the rational and dispassionate formulation of moral principles, rather than emotion. But Greene does not use the terms 'deontology' or 'consequentialism' in their ordinary sense. Instead, he characterises deontology and consequentialism functionally, in terms of two

---

[151] Greene, "Secret joke," 35–65.
[152] Ibid., 37.
[153] Ibid.
[154] Ibid.
[155] Christopher Meyers, "Brains, trolleys, and intuitions: Defending deontology from the Greene/Singer argument." *Philosophical Psychology* 28, no. 4 (2015): 466-486, at 468.

distinct kinds of moral judgements.[156] Characteristically deontological judgements, according to Greene, are those that place an emphasis on the intrinsic rightness or wrongness of actions, irrespective of the consequences of those actions.[157] And characteristically consequentialist judgements, by contrast, are those that place emphasis only on the overall outcome of actions.[158] Put otherwise, consequentialist judgments have the form 'X is wrong (or right) because it will bring about bad (or good) consequences.' And deontological judgements have the form 'X is just *wrong* (or just *right*), no matter what.'

In everyday contexts, the term 'emotion' is associated with feelings, and the term 'cognition' is associated with thinking. But Greene introduces a novel definition of the terms cognition and emotion, in terms of two distinct kinds of psychological process.[159] He defines emotion in terms of psychological processes that are fast, automatic, not necessarily conscious; and which have corresponding behavioural and physiological effects.[160] The neural correlates of emotion, according to Greene, are the amygdala, posterior cingulate cortex, medial prefrontal cortex, medial surfaces of the frontal and parietal lobes.[161] And he defines cognition in terms of psychological processes that are slow, deliberative, neutral, and which do not trigger automatic behavioural responses. Cognition, according to Greene, is responsible for "reasoning, planning, manipulating information in the working memory, controlling impulses, and higher executive function"; and has neural correlates in the dorsolateral surfaces of the prefrontal cortex, and parietal lobes.[162]

A common criticism of Greene's theory is that cognition and emotion are not discrete

[156] Greene, "Secret joke," 37.
[157] Ibid., 39.
[158] Ibid.
[159] Ibid., 40.
[160] Ibid., 41.
[161] Ibid., 40.
[162] Ibid.

kinds of mental process.[163] Loader's claim that people can have "thoughtful feelings and passionate thoughts" captures the core of the objection—that cognition and emotion are not discrete kinds of mental process.[164] In my view, this objection is strictly speaking correct, but it is uncharitable to Greene's intention in distinguishing between cognition and emotion. Greene's aim is not to assert that cognition and emotion are sharply separable psychological processes. Rather, his aim is show that deontological and consequentialist moral judgements seem to be underpinned by two importantly different patterns of moral thinking.[165] His distinction between cognition and emotion is an expository device, designed to pragmatically demarcate two aspects of human psychology.

Having clarified Greene's terminology, we are now in a position to return to his claim that deontology is generated by fast, automatic, emotional processes; whereas consequentialism is generated by slow, deliberate, neutral, and cognitive processes.[166] To empirically test this idea, Greene designed and ran a large number of experiments to determine the psychological processes that occur while individuals engage in moral judgement.[167] The general design of these experiments was to isolate deontological judgments from consequentialist judgements, by asking subjects hypothetical moral dilemmas while recording brain-imaging and reaction time data.[168] To get a sense of the evidence that Greene provides in support of his theory, it is worth describing one such experiment.

In the classic 'trolley problem' dilemma, a runaway trolley is hurtling towards five individuals tied to a train-track, who will all die if the trolley hits them.[169] The agent is given the choice to pull a lever that will divert the trolley to a side-track, thereby saving the five individuals—

---

[163] Rob Canton, "Crime, punishment and the moral emotions: Righteous minds and their attitudes towards punishment." *Punishment & Society* 17, no. 1 (2015): 54-72, at 58.
[164] Ian Loader, "Playing with fire? Democracy and the emotions of crime and punishment." *Emotions, Crime and Justice* (2011): 347-362, at 351.
[165] Greene, "Secret joke," 116.
[166] Meyers. "Brains, trolleys," 466-486.
[167] Meyers, "Brains, trolleys," 469.
[168] Greene, "Secret joke," 41–43.
[169] Ibid.

but on the side-track, a sixth individual is tied down, who, if the lever is pulled, will be hit by the trolley and die. In a modified version of the trolley problem, known as the 'footbridge problem,' instead of a side-track and a lever, there is a large person standing on a footbridge over the train-tracks.[170] If the large person is pushed off the footbridge and onto the tracks in front of the trolley, the trolley will run into the large person's body, and stop. The five individuals will be saved, but the large person will be killed. In this case, the moral dilemma is between two options: (1) push the large person to their death to save five lives, or (2) don't push the large person to their death, which would result in five deaths. To test his theory of moral judgement, Greene hypothesised that subjects who chose option (1) when presented with the footbridge problem would tend to show more brain activity in regions associated with cognition, because they would necessarily be making a consequentialist judgement.[171] The fundamental rationale behind pushing the large person off the bridge is that one death is better than five deaths—a cost-benefit analysis about consequences.[172] Subjects who choose option (2) in the footbridge problem, by contrast, would tend to show more brain activity in regions associated with emotion, because they would necessarily be making a deontological judgement.[173] Choosing option (2) shows an aversion to causing the death of the large person, despite an overall worse outcome (five deaths instead of one).[174] Greene also hypothesised that reaction times for individuals who gave the consequentialist response would be slower than those who gave the deontological response, because the controlled, cognitive calculation that one death is better than five deaths would take time to override the automatic emotional aversion to pushing the fat man off the bridge.[175] Both of Greene's hypotheses were confirmed, suggesting a strong correlation between deontology and emotion; and

---

[170] Ibid.
[171] Meyers, "Brains, trolleys, 469.
[172] Joshua Greene et al., "Cognitive load selectively interferes with utilitarian moral judgment." *Cognition* 107, no. 3 (2008): 1144-1154.
[173] Ibid., 44.
[174] Meyers, "Brains, trolleys," 469.
[175] Greene et al., "Cognitive load," 1149.

between consequentialism and cognition.[176]

To establish a causal relationship between cognition and consequentialism, in addition to a correlation, Greene et al., instructed test-subjects to monitor a scrolling string of random numbers while providing a response to the footbridge dilemma.[177] This modification was designed to impose a small load on the working-memory, leaving less cognitive resources for consequentialist reasoning. It was hypothesised that individuals who responded in a consequentialist way to the footbridge case would have slower response-times while under cognitive-loading than individuals who responded in a consequentialist way, but were not cognitively loaded. This prediction was confirmed, providing further evidence of the connection between cognition and consequentialism.[178] Greene also hypothesised that cognitive-loading would have no significant effect on the response-times of individuals who provided the deontological response to the footbridge case, on the grounds that deontological judgements are inherently emotional, rather than cognitive. Again, this hypothesis was confirmed.[179]

To clarify and lend support to Greene's theory, it is worth addressing the main objection it faces.[180] Namely, that consequentialist judgments are not solely the product of cognitive processes, and that deontological judgements are not exclusively emotion-driven.[181] Greene acknowledges that emotion plays a role in consequentialist judgement, but he notes that the emotions which figure in deontology function like an 'alarm bell', whereas the emotions that figure in consequentialism function to inform the process of rational deliberation:[182]

> "The sorts of emotions hypothesised to be involved [in consequentialism] say, 'Such-and-such matters this much. Factor it in.' In contrast, the emotions hypothesised to drive deontological judgment are far less subtle. They are . . . alarm signals that issue

[176] Meyers, "Brains, trolleys," 469.
[177] Greene et al., "Cognitive load," 1145.
[178] Ibid., 1152–1154.
[179] Ibid.
[180] Myers, "Brains, trolleys," 466.
[181] Greene, "Secret joke," 64–65.
[182] Ibid.

> simple commands: 'Don't do it!' or 'Must do it!'"[183]

Greene also acknowledges that cognitive processes may play in deontological judgements:

> "One could, in principle, make a characteristically deontological judgment by thinking explicitly about the categorical imperative and whether the action in question is based on a maxim that could serve as a universal law."[184]

But he proposes instead that deontological judgements are seldom performed in this way, and that deontology is instead "affective at its core."[185] Hence, Greene circumvents the criticism that consequentialist judgments are not solely the product of cognitive processes, and that deontological judgements are not exclusively emotion-driven.

4.2. The Link Between Moral Judgement and Punitive Decision-making

For present purposes, the relevant aspect of Greene's theory of moral judgement is his claim that characteristically deontological judgements are generated by emotion. This claim is commensurate with the psychological evidence about punitive decision-making that I outlined in §2.2, in that the retributive intuitions that motivated punitive judgement were driven by strong negative emotions such as moral outrage.[186]

Greene's explanation for why deontological judgment evolved also corresponds with the evolutionary story about human punishment that I described in §3.5.1–3.5.4. He explains that the emotions that generate deontological reasoning evolved for the biological function of motivating individuals to act in ways that "help spread [their] genes within a social context"—a direct link to the maintenance of cooperation in ancestral human life.[187] Moreover, Greene also explains that emotion was selected as the mechanism for deontological judgement for its capacity to provide

---

[183] Ibid., 64.
[184] Ibid., 65.
[185] Ibid.
[186] Carlsmith, Darley, and Robinson, "Why do we punish?" 295.
[187] Greene, "Secret joke," 59.

"very reliable, quick, and efficient responses to recurring situations, whereas reasoning is unreliable, slow and inefficient in such contexts."[188] This explanation chimes with the evolutionary explanation of the psychological mechanism for punishment that I gave in §2.3.

The fact that Greene's evolutionary explanation for the emotional basis of deontology dovetails with the function of the emotion-based psychological mechanism for punishment is not a coincidence. Punitive decision-making is increasingly being treated as a species of moral judgment.[189] And the fact that punitive decision-making is motivated by automatic, emotional reactions to non-cooperation, (and not by consequentialist considerations), indicates that punitive decision-making is species of what Greene refers to as 'characteristically deontological judgments.' That is, judgements that place an emphasis on the intrinsic rightness or wrongness of actions irrespective of the consequences of those actions.[190]

4.3. Deontological Moral Philosophy as Post Hoc Rationalisation

In this section, I will introduce and evaluate Greene's explanation for the existence of deontological moral philosophy. Although his account applies to deontological philosophy in general, I will focus on its relation to retributivism. By retributivism, I mean the super-category of philosophical theories that purport to justify punishment by appealing to the notion that guilty wrongdoers intrinsically deserve to be punished. Greene's explanation for the existence of deontological moral philosophy is derived from his theory that characteristically deontological judgements are driven by fast, automatic emotions and that these emotions evolved in humans for a selective function (see §4.1–4.2).

Humans are a highly intelligent species, with a brain that continuously tries to make sense of the world. A large body of studies indicate, however, that humans, by virtue of their advanced

[188] Ibid., 60.
[189] Carlsmith and Darley, "Psychological aspects," 213–214.
[190] Greene, "Secret joke," 39.

cognitive capacities, are prone to unintentionally justifying and explaining their own impulsive behaviours and emotional dispositions, when no such justification or explanation exists.[191] In one study, subjects were induced to prefer a product via psychological priming, and it was found that they later explained their preference for completely unrelated reasons.[192] In a different study, male subjects unconsciously misattributed anxiety and physiological arousal for sexual attraction.[193] It was found that male subjects who interacted with an attractive female experimenter immediately after crossing a precarious bridge over a deep canyon (intended to be a frightening experience), were twice as likely to call her and to ask for date than subjects who were given time to rest and de-stress before the interaction.

Evidence of the human tendency to post-hoc rationalise has also been revealed in cases of people with abnormal mental conditions. Patients with memory disorders are prone to making up stories to cover over their memory deficits, with no awareness of doing so.[194] Individuals who are instructed to perform certain behaviours while under hypnosis will invent explanations for their behaviour even though it was caused by the hypnotist.[195] And split-brain patients, whose cerebral hemispheres are disconnected from neuronal communication, are known to invent strange associations between completely unrelated objects when each of their hemispheres are shown different stimuli.[196]

Although the human tendency to post hoc rationalise can only be made salient in experimental contexts, it is a widely-accepted fact that humans are prone to cognitively interpreting their own emotions and behaviours by unconsciously forming coherent explanations, especially

---

[191] Greene, "Secret joke," 61.
[192] Richard Nisbett and Timothy Wilson, "The halo effect: Evidence for unconscious alteration of judgments." *Journal of personality and social psychology* 35, no. 4 (1977): 250–256.
[193] Donald Dutton and Arthur Aron, "Some evidence for heightened sexual attraction under conditions of high anxiety." *Journal of personality and social psychology* 30, no. 4 (1974): 510–517.
[194] Donald Stuss et al., "An extraordinary form of confabulation." *Neurology* 28, no. 11 (1978): 1166-1166.
[195] George Estabrooks, "Hypnotism." (Dutton & Co: New York, 1943): 7-82.
[196] Michael Gazzaniga et al., "Plasticity in speech organization following commissurotomy." *Brain* 102, no. 4 (1979): 805-815.

when no such explanations exist.[197] By combining this psychological fact with the fact that humans are largely unaware of the extent to which the intuitions that dictate moral judgement are driven by automatic emotional reactions to wrongdoing, Greene infers that deontological moral philosophy, as it stands, is a post hoc rationalisation of the automatic emotions and intuitions that evolved in ancestral humans for a purely selective function.[198]

> "What should we expect from creatures who exhibit social and moral behaviour that is driven largely by intuitive emotional responses and who are prone to rationalisation of their behaviours? The answer, I believe, is deontological moral philosophy."[199]

In §3.5.4, I presented the empirically-supported view that negative emotion evolved as the mechanism for punishment in humans. This view coincided exactly with the psychological evidence about punitive decision-making that I examined in §2.2. Punishment was solely motivated by automatic retributive intuitions, which were driven by strong negative emotions in reaction to wrongdoing, such as moral outrage. And §4.1, I examined Greene's theory and empirical evidence in support of the view that moral judgement—specifically, characteristically deontological judgement—is to a large extent motivated by motivated by fast, automatic, emotions. To explain this, Greene appeals to evolution by natural selection.[200] He argues that emotions motivated ancestral humans to perform behaviours that were selectively advantageous, such as altruistic cooperation and punishment, and that this ancient psychology persists in the modern-day humans.[201] This explanation accords with my account of the selection function of emotion in §3.6.

Greene's explanation for the existence of deontological moral philosophy amounts to the claim that humans extend their tendency to post hoc rationalise into the academic realm—by constructing elaborate, linguistically-formulated rationalisations in the form of deontological moral

---

[197] Greene, "Secret joke," 62.
[198] Ibid.
[199] Ibid., 61–62.
[200] Ibid., 59.
[201] Ibid.

philosophy. These rationalisations entail that actions are morally right or wrong just in virtue of their internal features, and that right actions have a reason-providing authority that makes them morally required. In Greene's view, however, the content of these deontological theories is simply post hoc rationalisation of the emotionally-driven intuitions that influence human moral judgments.[202]

> "Talk about rights, respect for persons, and reasons we can share are natural attempts to explain, in 'cognitive' terms, what we feel when we find ourselves having emotionally driven intuitions. . . Although these explanations are inevitably incomplete, there seems to be 'something deeply right' about them, because they give voice to powerful moral emotions."[203]

In Greene's view, the reason why so much of human moral reasoning is characteristically deontological—'X is just *wrong* (or just *right*), no matter what'—is that much of human moral intuition is driven by evolved emotional impulses.[204] In other words, when humans experience a strong and automatic aversion to certain actions, such as stealing, or cannibalism; this feeling is the biological, alarm-like emotional mechanism that evolved for a selective function coming into effect.[205] Given that humans are naturally inclined to interpret the world around them in a cognitive way, however, this evolved emotional mechanism consciously registers as an immediate, passionate urge for or against certain actions.[206] In Greene's view, these characteristically deontological moral intuitions are the cognitive interpretation of the visceral, emotions mechanism that evolved in ancestral humans.

Some psychologists, such as Jonathon Haidt, go so far as to suggest that moral reasoning as a whole is likely to be little more than a rationalisation of our intuitive, emotional responses—

---

[202] Ibid., 63.
[203] Ibid., 74.
[204] Ibid., 64.
[205] Ibid.
[206] Ibid., 59.

as Haidt puts it, 'the emotional dog wags the rational tail.'[207] But Greene's proposition is not as strong as this.[208] Greene's claim, in essence, is that deontological moral philosophy is nothing more than a sophisticated confabulation of the emotional impulses that evolved in ancestral humans for the function of motivating cooperation, punishment, and other social behaviours.

4.4. The Deontological Challenge

Before I explain how Greene's account of deontological moral philosophy relates to retributivism, however, it is worth examining a counterargument to his thesis. Greene's position on deontological moral philosophy, overall, is that it is nothing more than a sophisticated confabulation of the emotional mechanisms that evolved in ancestral humans for the function of motivating cooperation, punishment, and other social forms of behaviour. It is worth dwelling for a moment on the ambition of Greene's claim. His conclusion threatens to explain away the entirety of deontological moral philosophy: a longstanding theory and esteemed normative theory. Excessive ambition is sometimes a worrying sign in philosophy—but the strength of Greene's argument lies in his robust evolutionary explanation for the emotional source of deontology, and the large body of empirical evidence that he uses in support. Although I am inclined to agree with Greene, a compelling counterargument to his view needs to be addressed.

Traditionally, deontological moral philosophy is anchored in rationalism.[209] Roughly speaking, rationalism is the idea that reason and rationality are superior to the senses and sentimentality across all domains of philosophical inquiry.[210] In the domain of ethics, rationalist approaches to morality generally hold that moral judgement is based first and foremost in belief, rather than in emotion.[211] The overarching goal of rationalist deontological moral philosophy,

---

[207] Haidt, "The emotional dog," 814–834.
[208] Greene, "Secret joke," 63.
[209] Mark Timmons, "2.2 Toward a Sentimentalist Deontology." *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* 3 (2008): 93–104, at 98.
[210] Ibid.
[211] Timmons, "Towards," 102.

exemplified in the works of Kant, is to formulate an elegant and comprehensive set of fundamental moral principles, grounded in the intrinsic value of human beings, with the aim of determining right and wrong once-and-for-all.[212] For this reason, rationalist deontology is sometimes associated with moral realism—the idea that there is an independent realm of moral facts that determine the truth of moral beliefs.[213] As Greene writes,

> "[M]orals . . . stand alone like mathematical theorems, independent of the messy world of psychology. That is the deontological dream."[214]

Greene's evidence for the role of emotion in deontology, and his evolutionary explanation for why moral emotions evolved, presents rationalist deontologists with a significant explanatory burden. Why should we believe that there exists a deep, independent moral truth, or set of perfect moral principles, when a significant portion of human moral intuitions are simply driven by emotions, the existence of which can easily be explained in terms of evolution by natural selection?[215]

Although Greene's account presents a major problem for rationalist deontologists—certain contemporary deontologists seem able to resist his challenge. Constructivist deontologists, such as T. M. Scanlon, do not ground their conception of moral truth in the idea of a perfect set of independent, rationally attainable moral principles.[216] Rather, they attempt to *construct* moral truth, via a process of rationally reflecting on the unrefined moral intuitions and commitments that humans share. Examples of these principles include a respect for persons, and an aversion to treating people as mere objects.[217] The aim of constructivist deontology, broadly speaking, is to

---

[212] Joshua Greene, "2.3 Reply to Mikhail and Timmons." *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* 3 (2008): 105–115
[213] Timmons, "Towards," 97.
[214] Greene, "Reply," 115.
[215] Greene, "Secret joke," 76.
[216] Timothy Scanlon, *What we owe each other.* (Cambridge, MA: Harvard University Press, 1998), 153–168.
[217] Timmons, "Towards," 96.

refine these common intuitions and commitments into a set of moral principles, which will serve as a standard by which moral judgements can be justified.[218] Ostensibly, constructivist deontologists are able to resist Greene's claim that deontology is post hoc rationalisation, by admitting the evolutionary origins of characteristically deontological moral intuitions, whilst also constructing a deontological system of ethics in which moral principles have a reason-providing authority that makes them morally required.[219]

Greene's reply to this, however, is that the unrefined intuitions that *go into* the constructivist's rational reflection process will invariably include emotionally-generated characteristically deontological judgements.[220] So any set of moral principles that *come out* of the reflective process will necessarily contain post hoc rationalisation of emotions that evolved for a selective function. And the putative 'moral truth' found in the constructed set of principles would instead merely "reflect arbitrary features of our evolutionary history."[221]

Greene grants that it might be possible (at least in theory) to remove, or 'filter' out characteristically deontological judgments during the rational reflection process, thereby eliminating all emotionally-driven intuition from the constructivist's final set of moral principles.[222] The resulting ethical system would certainly not contain any post hoc rationalization. But what would such an ethical system look like? In Greene's view, it would look like the ethical system envisaged by the economist John Harsanyi in 1953.[223] Drawing from the economic theory of rational decision-making under uncertainty, Harsanyi proposed the idea of 'social welfare utility function,' whereby ethical decisions could be made on the basis of an impartial calculation about the common morality of all human beings.[224] On Harsanyi's system, all ethical decisions would be

---

[218] Ibid., 100.
[219] Timmons, "Towards," 101.
[220] Greene, "Reply," 116.
[221] Ibid.
[222] Ibid., 117.
[223] Ibid.
[224] John Harsanyi, "Cardinal utility in welfare economics and in the theory of risk-taking." *Journal of Political Economy* 61, no. 5 (1953): 434-435.

made from the perspective of an impartially-situated individual, whose interests and personal biases were totally removed by "being put in a situation of complete uncertainty about their true identity."[225] According to Greene, an ethical system such as Harsanyi's would indeed eliminate post hoc rationalisation. But it would amount to little more than a "utilitarian philosophy mounted upon a would-be deontological foundation"—that is, a calculative method for weighing up right and wrong, based on a single moral principle.[226] According to Greene, this kind of ethical system, though free of characteristically deontological judgments, would be more consequentialist than deontological. And, therefore, such a system would not serve as a counterexample to the claim that all deontological moral philosophy involves the post hoc rationalisation of emotion.[227]

4.5. Retributivism as Post Hoc Rationalisation

Having explicated and defended Greene's explanation for the existence of deontological moral philosophy, I will now employ his account to explain the existence of retributivism.

In §4.1, I raised the question: where does the philosophical theory of retributivism fit into the overarching evolutionary story about the biological function and psychological mechanism for punishment in humans? As I explained in §2.1, the term retributivism applies to a wide category of philosophical theories that attempt to justify punishment by claiming that guilty wrongdoers morally deserve proportional punishment, irrespective of any beneficial consequences that punishment brings about. The central principle of retributivism is 'desert'—the idea that punishment is the morally required response to certain kinds of wrongful acts.

Retributivism is often referred to as a 'deontological' theory of punishment, due to the fact that retributivists consider punishment intrinsically valuable, despite the good consequences that it might bring about. Given that retributivism is often described as a deontological theory of

---

[225] Peter Hammond, "Harsanyi's utilitarian theorem: A simpler proof and some ethical connotations." In *Rational Interaction* Springer Berlin: Heidelberg (1992): 305-319.
[226] Greene, "Reply," 116.
[227] Ibid.

punishment, it is unsurprising that its contents correspond with what Greene calls characteristically deontological judgments, which have the form 'X is just *wrong* (or just *right*), no matter what.' On retributivism, guilty wrongdoers just *deserve* proportional punishment—end of story. Yet according to Greene's theory of the psychological processes that underpin moral judgment, characteristically deontological judgements are generated by fast, automatic, emotional reactions to wrongdoing, which evolved for the function of sustaining social behaviour in ancestral human groups.

Ordinarily, retributivism is treated as a philosophical theory that emerged from rational reflection about the fundamental moral basis of punishment. But Greene's theory of moral judgement seems to suggest that retributivism (one of the oldest and most widely-endorsed theories regarding the justification of punishment) is driven merely by moral emotions that evolved for the function of maximising reproductive success in ancestral humans. Could this be true?

In chapter 3, I presented Bowles and Gintis' hypothesis for the evolution of punishment in ancestral humans, to explain why punitive decision-making in modern-day humans is immediately motivated by emotionally-driven retributive intuitions, and not by considerations about the benefits of punishment. In the later stages of the Pleistocene epoch, hunter-gatherers were subject to a great number of survival pressures, including resource scarcity, a volatile climate, and the risk of lethal intergroup conflict. Groups comprising individuals with a predisposition for helping one another, even when doing so came with personal costs, had a significant reproductive advantage. Cooperation was highly beneficial, but it also opened the possibility for cheaters and norm-violators to exploit its benefits. To solve this problem, and to sustain within-group cooperation, ancestral humans evolved a mechanism for modifying the behaviour of non-cooperative individuals. This mechanism, punishment, operated by motivating individuals to cause non-cooperators harm; which conditioned those individuals against non-cooperative behaviour in the future.

In this chapter, I explicated and defended Greene's explanation for the existence of deontological moral philosophy. According to Greene, deontological moral philosophy is nothing

more than elaborate, systematic, verbally-expressed post hoc rationalisation of the moral intuitions and emotions that evolved for the function of maximising reproductive success in ancestral humans. By connecting my earlier evolutionary exposition of punishment with Greene's explanation for the existence of deontological moral philosophy, it is now possible construct a tentative explanation for the existence of the philosophy theory of retributivism.

Punishment, which evolved in humans for a selective function, is triggered by an emotional mechanism. This mechanism evolved because it was automatic, effective and reliable. Rather than performing slow, cognitive cost-benefit analyses about the long-term benefits of punishment, ancestral humans simply experienced automatic and visceral feelings of moral outrage in response instances of free-riding and norm-violation. The intensity of these negative emotions evolved to track (a) the blameworthiness of non-cooperative individuals, and (b) the magnitude of harm caused by the non-cooperative behaviour, because it was selectively disadvantageous to deliver punishment that was overly-severe, or overly-lenient. This ancient punitive psychology persists in modern-day humans (see §2.2).

To a modern reader, the above-mentioned evolutionary story is quite intelligible. But none of the above was evident to the philosophers who were formulating the first justificatory theories of punishment, such as Kant, or Hegel. Instead, these philosophers were simply aware of a deeply-held intuition—in their own psychology, and in that of others—that punishment was the appropriate response to wrongdoing. They engaged in a cognitive appraisal of this intuition, without knowing that they were interpreting an evolutionarily ancient psychological mechanism. They chose the concept of 'desert', or 'deservingness' to represent the irrepressible psychological impulse to punish wrongdoers that they were cognizant of. They chose the concept of 'proportionality' to represent the strong psychological sense that punishing too severely or too leniently was wrong. And they systematised these concepts into the form of a justificatory theory: the idea that guilty wrongdoers intrinsically deserve to be proportionally punished. If this explanatory story is correct, then retributivism is not, first and foremost, a justificatory theory of

punishment. Rather, retributivism is an elegant and systematic post hoc rationalisation of the emotionally-driven psychological mechanism that evolved to motivate punishment in groups of ancestral humans.

The idea that retributivism is post hoc rationalisation raises several important raises questions. First, and most obvious, if retributivism is merely a post hoc rationalisation, then those who espouse retributivist theories of punishment are largely oblivious to both the psychological predisposition that motivates their position, and to the evolutionary origins of that predisposition. If the present view of retributivism is correct, should those who endorse retributivism continue to do so? According to Greene, and also Singer, the answer to this question is no—or, at least, not to the same extent. Greene claims that understanding the roots of one's evolutionary impulses should reduce their moral authority.[228] And according to Singer, "advances in our [evolutionary] understanding of ethics . . . gives us grounds for being less respectful of them."[229]

A different question that arises pertains to the philosophical literature on retributivism. For centuries, philosophers have provided diverse and illuminating formulations of retributivism, and penetrating analyses of the concepts of desert and proportionality. If retributivism is a merely a post hoc rationalisation, do these contributions lose their theoretical utility? Should retributivist doctrines continue to be taught in undergraduate law and philosophy courses?

From these theoretical questions, further questions arise regarding practical matters, such judicial sentencing, punitive legislation, and the general administration of criminal justice institutions.[230] Justificatory theories of punishment have real, downstream consequences for prison systems, sentencing regimes, and incarcerations rates.[231] If retributivism is, in fact, simply a vestigial relic from the psychology of our hunter-gatherer ancestors, then why should it have a place among the jurisprudential principles that govern contemporary penal institutions?

---

[228] Greene, "Secret joke," 76.
[229] Peter Singer, "Ethics and intuitions." *The Journal of Ethics* 9, no. 3-4 (2005): 331-352, at 349.
[230] Hoffman, *Punisher's brain*, 337.
[231] Ibid.

In my view, each of the above-mentioned questions hinge on a broader problem, relating to the theoretical status of retributivism. Is it appropriate to continue to regard retributivism a justificatory theory of punishment? If retributivism is merely the post hoc rationalisation of an evolutionarily ancient psychological mechanism, then to my mind, retributivism should no longer be regarded as a justificatory theory of punishment. If the above conclusion is correct, then making a justificatory appeal to retributivism when delivering punishment is essentially the same—to put it crudely—as claiming: 'my hunter-gatherer ancestors acted in this way, so I can act in this way too.' Because of this, I am of the opinion that the theoretical status of retributivism *qua* justificatory theory should change. But what should it change *to*?

Describing retributivism as 'post hoc rationalisation' is strictly speaking accurate, but it is also a somewhat pejorative label. In my view, a slightly less precise, but more charitable way to think about retributivism is from an evolutionary standpoint. Retributivism is the linguistic expression of an evolutionarily ancient psychological mechanism for punishment. Regarding retributivism in this way has three merits. First, it preserves the theoretical accomplishments of punishment theorists in formulating precise versions of retributivism, and in elucidating its core concepts. Second, it neutralises the idea that retributivism has any real justificatory power. And third, it explicitly acknowledges the connection between retributivism and our hunter-gatherer ancestry.

5. Conclusion

The first aim of this paper was to provide an evolutionary explanation for why punitive decision-making in modern humans is motivated by emotionally-driven retributive intuitions. This aim was met in the form of an empirically-informed evolutionary exposition of the origin, phylogenetic distribution, selective function, and development of punishment, which culminated in the psychological mechanism for punishment: negative emotion. Anger, indignation, and outrage form the motivational mechanism for punishment in humans, and this explains why

humans do not consider the consequences of punishment when making practical punitive decisions.

On a more methodological note, the structure of my evolutionary exposition of human punishment—based on a modified version of Tinbergen's framework—was advantageous, for three main reasons. First, it provided a clear demarcation between important conceptual distinctions in biology, such as the ontogeny–phylogeny distinction, and the mechanism–function distinction. Second, it divided up the explanatory labour in such a way that turned an otherwise opaque explanatory project into a relatively straightforward task. And third, it allowed for a diverse range of interdisciplinary findings to be coherently synthesised.

The second aim of this paper was to re-evaluate retributivism from an evolutionary perspective. This aim was met by applying Greene's ground-breaking explanation for the existence of deontological moral philosophy to the evolutionary exposition of human punishment that I provided in chapter 3. The finding from this re-evaluation was that retributivism is a post hoc rationalisation of the psychological mechanism that evolved to motivate punishment in ancestral humans. Undoubtedly, this finding has serious theoretical and practical implications—of which I was only able to offer a brief sketch. In my view, however, one implication is particularly conspicuous. Retributivism should no longer be accorded the theoretical status of a justificatory theory of punishment. To continue to treat retributivism in this way is to wilfully ignore the human proclivity for cognitively interpreting the emotion-driven mechanisms that were built into our psychology by natural selection.

It is important to clarify however, that although I do not think retributivism should retain a reason-providing authority, I am certainly not of the opinion that it should be disparaged or cast aside in any way. Instead, I think that retributivism should simply be acknowledged for what it actually *is*—a refined linguistic expression of an ancient punitive predisposition. It is quite a remarkable fact, after all, that humans are able to philosophically reflect on, and verbally articulate

the conscious registration of the emotionally-driven psychological mechanism that evolved to secure the survival and reproductive success of our hunter-gatherer ancestors.

References

Allen, Ronald, and Brian Leiter. "Naturalized epistemology and the law of evidence." *Virginia Law Review* (2001): 1491–1550.

Barclay, Pat, and Mark Van Vugt. "The evolutionary psychology of human prosociality: Adaptations, byproducts, and mistakes." in *The Oxford Handbook of Prosocial Behaviour* (Oxford Library of Psychology, 2015), 37-60.

Bedau, Hugo and Erin Kelly. "Punishment," *The Stanford Encyclopedia of Philosophy* (Fall 2015), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2015/entries/punishment/>.

Bentham, Jeremy. *An Introduction to the Principles of Morals and Legislation (Chapters I–V)*. Blackwell Publishing Ltd, [1789] 1972.

Bowles, Samuel, and Herbert Gintis. *A Cooperative Species: Human Reciprocity and its Evolution*. (Princeton University Press, 2013).

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter Richerson. "The evolution of altruistic punishment." *Proceedings of the National Academy of Sciences* 100, no. 6 (2003): 3531-3535.

Calcott, Brett. "Why the proximate-ultimate distinction is misleading, and why it matters for understanding the evolution of cooperation." *Cooperation and its evolution* (2013): 249–263.

Canton, Rob. "Crime, punishment and the moral emotions: Righteous minds and their attitudes towards punishment." *Punishment & Society* 17, no. 1 (2015): 54-72.

Carlsmith, Kevin, and John Darley. "Psychological aspects of retributive justice." *Advances in Experimental Social Psychology* 40 (2008): 193–236.

Carlsmith, Kevin, John Darley, and Paul Robinson. "Why do we punish? Deterrence and just deserts as motives for punishment." *Journal of personality and social psychology* 83, no. 2 (2002): 284–299.

Carlsmith, Kevin. "The roles of retribution and utility in determining punishment." *Journal of Experimental Social Psychology* 42, no. 4 (2006): 437–451.

Clutton-Brock, Tim, and Geoffrey Parker. "Punishment in animal societies." *Nature* 373, no. 6511 (1995): 209–216.

Cushman, Fiery. "Punishment in humans: From intuitions to institutions." *Philosophy Compass* 10, no. 2 (2015): 117–133.

Darley, John, Kevin Carlsmith, and Paul Robinson. "Incapacitation and just deserts as motives for punishment." *Law and Human behavior* 24, no. 6 (2000): 659–683.

Davies, Nick, John Krebs and Stuart West. *An Introduction to Behavioural Ecology, 4th ed.,* (Wiley-Blackwell: Chichester UK, 2012).

De Quervain, Dominique, Urs Fischbacher, Valerie Treyer, and Melanie Schellhammer. "The neural basis of altruistic punishment." *Science* 305, no. 5688 (2004): 1254–1258.

Dewey, John. "The Influence of Darwin on Philosophy." in *The Influence of Darwin on Philosophy and Other Essays.* New York: Henry Holt and Company (1910): 1–19.

Dunbar, Robin. "The social brain hypothesis." *Brain* 9, no. 10 (1998): 178-190.

Dutton, Donald, and Arthur Aron. "Some evidence for heightened sexual attraction under conditions of high anxiety." *Journal of personality and social psychology* 30, no. 4 (1974): 510–517.

Estabrooks, George. "Hypnotism." (Dutton & Co: New York, 1943): 7-82.

Exodus 21: 23–25, *The Holy Bible: King James Version.* Texas: National Publishing Company (2000).

Fehr, Ernst, and Simon Gächter. "Altruistic punishment in humans." *Nature* 415, no. 6868 (2002): 137-140.

Francis, Richard. "Causes, proximate and ultimate." *Biology and Philosophy* 5, no. 4 (1990): 401-415.

Frank, Robert. *Passions within reason: the strategic role of the emotions.* WW Norton & Co, 1988.

Gazzaniga, Michael, Bruce Volpe, Charlotte Smylie, Donald Wilson, And Joseph Le Doux. "Plasticity in speech organisation following commissurotomy." *Brain* 102, no. 4 (1979): 805-815.

Gintis, Herbert. "Gene–culture coevolution and the nature of human sociality." *Philosophical Transactions of the Royal Society B: Biological Sciences* 366, no. 1566 (2011): 878–888.

Godfrey-Smith, Peter. "A modern history theory of functions." *Noûs* 28, no. 3 (1994): 344–362.

——— *Philosophy of biology*. (Princeton University Press, 2013).

Golash, Deirdre. *The Case Against Punishment*. (New York: New York University Press, 2005), 60–71.

Gough, Ian, Garry Runciman, Ruth Mace, Geoffrey Hodgson, and Michael Rustin. "Darwinian evolutionary theory and the social sciences." *Twenty-First Century Society* 3, no. 1 (2008): 65-86.

Greene, Joshua, Sylvia Morelli, Kelly Lowenberg, Leigh Nystrom, and Jonathan Cohen. "Cognitive load selectively interferes with utilitarian moral judgment." *Cognition* 107, no. 3 (2008): 1144-1154.

Greene, Joshua. "2.3 Reply to Mikhail and Timmons." *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* 3 (2008): 105–115

——— "The secret joke of Kant's soul." *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* 3 (2008): 35–79.

Haidt, Jonathan. "The emotional dog and its rational tail: a social intuitionist approach to moral judgment." *Psychological review* 108, no. 4 (2001): 814–834.

Haig, David. "Proximate and ultimate causes: how come? and what for?" *Biology & Philosophy* 28, no. 5 (2013): 781–786.

Hammond, Peter. "Harsanyi's utilitarian theorem: A simpler proof and some ethical connotations." in *Rational Interaction* Springer Berlin: Heidelberg (1992): 305-319.

Harsanyi, John. "Cardinal utility in welfare economics and in the theory of risk-taking." *Journal of Political Economy* 61, no. 5 (1953): 434-435.

Hart, H. L. A. "Prolegomenon to the Principles of Punishment." *Proceedings of the Aristotelian Society* 60 (1959–1960): 1–26.

Hegel, Georg W. F. *Elements of the Philosophy of Right*. Trans. N. B. Nisbet, Ed. Allen Wood. (Cambridge: Cambridge University Press, [1821] 1991). 119–131.

Hoffman, Morris. *The Punisher's Brain: The Evolution of Judge and Jury*. Cambridge University Press, 2014.

Hoffman, Morris, and Timothy Goldsmith. "The biological roots of punishment." *Ohio State Journal of Criminal Law*. 1 (2003): 627–641, at 627.

Kahneman, Daniel, David Schkade, and Cass Sunstein. "Shared outrage and erratic awards: The psychology of punitive damages." *Journal of Risk and Uncertainty* 16, no. 1 (1998): 49–86.

Kahneman, Daniel. *Thinking, fast and slow*. (Macmillan, 2011), 20–23.

Kant, Immanuel. *The Metaphysics of Morals*, M. Gregor (trans.), (New York: Cambridge University Press [1797] 1991).

Loader, Ian. "Playing with fire? Democracy and the emotions of crime and punishment." *Emotions, Crime and Justice* (2011): 347-362.

Mayr, Ernst. "Cause and effect in biology." *Science* 134, no. 3489 (1961): 1501-1506.

Meyers, Christopher. "Brains, trolleys, and intuitions: Defending deontology from the Greene/Singer argument." *Philosophical Psychology* 28, no. 4 (2015): 466-486.

Murphy, Jeffrey, and Jean Hampton. *Forgiveness and Mercy*. Cambridge: Cambridge University Press, 1988.

Nakao, Hisashi and Edouard Machery. "The evolution of punishment." *Biology and Philosophy* 27, no. 6 (2012): 833–850.

Nisbett, Richard, and Timothy Wilson, "The halo effect: Evidence for unconscious alteration of judgments." *Journal of personality and social psychology* 35, no. 4 (1977): 250–256.

Rawls, John. "Two Concepts of Rules." *Philosophical Review* 64 (1955): 3–32.

Scanlon, Timothy. *What we owe each other.* (Cambridge, MA: Harvard University Press, 1998).

Sherbrooke, Wade, and George Middendorf. "Blood-squirting variability in horned lizards (Phrynosoma)." *Copeia* 2001, no. 4 (2001): 1114–1122.

Singer, Peter. "Ethics and intuitions." *The Journal of Ethics* 9, no. 3-4 (2005): 331-352.

Sripada, Chandra. "Punishment and the strategic structure of moral systems." *Biology and philosophy* 20, no. 4 (2005): 767–789.

Stuss, Donald, Michael Alexander, Aubrey Lieberman, and Harvey Levine. "An extraordinary form of confabulation." *Neurology* 28, no. 11 (1978): 1166-1166.

Timmons, Mark. "2.2 Toward a Sentimentalist Deontology." *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* 3 (2008): 93–104.

Tinbergen, Nikolaas. "On aims and methods of ethology." *Zeitschrift für Tierpsychologie,* 20, (1963): 410–433.

Treadway, Michael, Joshua Buckholtz, Justin Martin, Katharine Jan, Christopher Asplund, Weiner, Bernard, Sandra Graham, and Christine Reyna. "Corticolimbic gating of emotion-driven punishment." *Nature Neuroscience* 17, no. 9 (2014): 1270–1275.

Trivers, Robert. "The evolution of reciprocal altruism." *The Quarterly review of biology* 46, no. 1 (1971): 35-57.

Walen, Alec. "Retributive Justice." *The Stanford Encyclopedia of Philosophy* (Winter 2016), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/justice-retributive/>.

Weiner, Bernard, Sandra Graham, and Christine Reyna. "An attributional examination of retributive versus utilitarian philosophies of punishment." *Social Justice Research* 10, no. 4 (1997): 431–452.

Wiegman, Isaac. "The Evolution of Retribution: Intuitions Undermined." *Pacific Philosophical Quarterly*, 98, (2017): 193–218.

Wood, David. "Punishment: consequentialism." *Philosophy Compass* 5, no. 6 (2010): 455–469.

———"Punishment: nonconsequentialism." *Philosophy Compass* 5, no. 6 (2010): 470–482.

Wright, Larry. "Functions." *The philosophical review* 82, no. 2 (1973): 139-168.

———*Teleological explanations: An etiological analysis of goals and functions*. (University of California Press, 1976).