RICE UNIVERSITY

Punishment by Agreement – A Contract-Based Argument for the Moral Permissibility of Punishment

by

Daniel Burkett

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

George Sher, Chair
Herbert S. Autrey Professor of Philosophy

Gwen Bradford
Associate Professor of Philosophy

(Deceased)

Baruch Brody
Professor Emeritus

John H. Zammito
John Antony Weir Professor of History

HOUSTON, TEXAS
16 November 2018
Abstract

Punishment necessarily involves the infliction of harm, and therefore requires us to treat wrongdoers in ways that it would clearly be wrong to treat others. Traditionally, this harm is justified on the basis of either (1) the Consequentialist claim that punishment maximises the good for the wider community, or (2) the Retributivist claim that wrongdoers simply deserve the harm they receive. However, both theories are subject to well-known objections. Specifically, there are a variety of cases in which each theory either punishes the innocent, or fails to punish the guilty. This leads some authors to the controversial conclusion that punishment is never morally justified. I propose an alternative solution: I argue that the moral permissibility of punishment should instead be based on the fact that each wrongdoer has agreed to her punishment. Specifically, I argue that it will be morally permissible to punish an individual $P$ under a particular punishment practice $X$ so long as—at some prior point—$P$ simultaneously (i) had good reason to agree to $X$, and (ii) gave some kind of agreement to $X$. I then argue that (i) will be met where an individual can expect to benefit from the implementation of a punishment practice, and—more controversially—that (ii) can be satisfied by the existence of hypothetical agreement. I reflect on the practical implications of this theory by applying it to a number of real-world crimes, and demonstrating how it can go about prescribing specific punishments for particular offences. I conclude by showing that the Contractarian theory is well-equipped to avoid the objections commonly levelled against traditional theories of punishment, and that—for this reason—it provides a promising solution to the problem of punishment.
Acknowledgements

My thanks goes—first and foremost—to Professor George Sher for his extensive advice and assistance throughout the thesis-writing process. His ability to take my most nebulous, ill-formed thoughts and crystallise them into coherent arguments lies behind almost every page of this dissertation. I am also indebted to Professors Gwen Bradford and John H. Zammito, whose invaluable feedback and critiques on the final draft of this thesis gave me the opportunity to refine—and anticipate several objections to—my core argument. While he passed away before seeing my final thesis, I am incredibly grateful for Professor Baruch Brody’s instrumental contributions very early on in the writing process. His formidable ability to identify the chinks in any philosophical armour helped me to see where the greatest weaknesses in my argument would lie, and where the most work would need to be done.

I am deeply appreciative of the wider Rice community for all they have done during my tenure as a graduate student. In particular, the unending support and collegiality of the Department of Philosophy has always made me feel as though I belong, despite being thousands of miles from home. I am especially grateful for the friendship of my classmates—most notably my cohort of Roma Hernández, Graham Valenta, Brandon Williams, and Peter Zuk—with whom my many philosophical conversations at Valhalla helped plant the seeds of this dissertation. My thanks also goes to my parents, without whose love, encouragement, and support I might never have considered pursuing graduate study in a foreign country.

My greatest thanks goes to my fiancée—and best friend—Holly. Her belief in my abilities has always far outstripped my own, and her unwavering support and encouragement was vital in helping me see this project through to completion. Her kindness, patience, and perpetual
willingness to engage in philosophical debate went above and beyond what any person could ever expect of another human being. Words will forever fail to convey the depth of my love and appreciation for her.
# Table of Contents

**Chapter 1: The Problem of Punishment**

- What is Punishment? ................................................................................................................... 2
- The Problem of Punishment ........................................................................................................ 7
- Concerns with the Consequentialist Solution to the Problem of Punishment .................................. 9
  - Act Utilitarianism .................................................................................................................. 10
  - Rule Utilitarianism .............................................................................................................. 14
  - Non-Utilitarian Consequentialism .................................................................................... 17
- Concerns with the Retributivist Solution to the Problem of Punishment ........................................ 18
  - Desert-Based Retributivism .............................................................................................. 18
  - Forfeiture-Based Retributivism ........................................................................................ 21
  - Fairness-Based Retributivism ........................................................................................... 25
- A New Solution to the Problem of Punishment ............................................................................. 29

**Chapter 2: Outlining a Contractarian Theory of Punishment** ................................................. 31

- “Contractarianism” ................................................................................................................... 32
- Distinguishing Contractarianism from Contractualism ............................................................ 37
- A Contractarian Theory of Punishment .................................................................................... 40

**Chapter 3: How We Have Good Reason to Agree to Punishment** ........................................... 48

- The Disutilities of Sanction, Restraint, and Collateral ............................................................. 49
- To Agree or Not to Agree ......................................................................................................... 52
- Guaranteed Benefit ................................................................................................................... 56
- A Lower Standard than Guaranteed Benefit? ........................................................................... 61
- Expected Benefit ....................................................................................................................... 64
- Having Good Reason to Agree to Punishment ......................................................................... 69

**Chapter 4: How We Agree to Punishment** .............................................................................. 74

- Express Agreement ................................................................................................................... 75
- Implied Agreement ................................................................................................................... 78
- Unanimity and Tacit Agreement .............................................................................................. 81
- Hypothetical Agreement ........................................................................................................... 88
- When We Agree to Punishment ................................................................................................. 97
Chapter 5: The Contractarian Theory of Punishment in Practice ................................. 102
  Consolidating the Contractarian Test............................................................................ 102
  Has Agreement Become Superfluous? .......................................................................... 105
  Is the Threshold of Unanimity Too High? .................................................................... 109
  The Theory in Practice................................................................................................. 114
  A Brief Reflection........................................................................................................... 127

Chapter 6: Problem Cases for a Contractarian Theory of Punishment ......................... 129
  The First Case: The Value Outlier.................................................................................. 130
  The Second Case: The Vigilante.................................................................................... 140
  The Third Case: The Foole............................................................................................ 143
  Conclusion on Problem Cases...................................................................................... 153

Chapter 7: A Solution to the Problem of Punishment...................................................... 155
  A Concern with the Practical Implementation of the Contractarian Theory.................. 155
  How Contractarianism Avoids Concerns with Traditional Theories of Punishment ....... 163
  Conclusion..................................................................................................................... 172

Bibliography..................................................................................................................... 174
Chapter 1: The Problem of Punishment

Many see punishment as an appropriate response to certain kinds of wrongdoing. We might disagree about the nature or the extent of punishment that is appropriate, but most of us endorse the idea that someone who commits a crime should suffer some kind of negative consequences for his or her actions. Despite this, the act of punishing an offender remains deeply problematic. This is because punishment necessarily involves the infliction of harm,\(^1\) and therefore requires us to treat wrongdoers in ways that it would clearly be wrong to treat others. What is needed, then, is an account of why certain kinds of harm are morally permissible when carried out in the name of punishment. This is the so-called “problem of punishment.”

According to David Boonin, punishment is never morally permissible.\(^2\) Boonin defends this claim by attempting to refute a wide range of arguments that purport to morally justify punishment. He notes, however, that this strategy does leave open the possibility that there is some undiscovered theory that may yet provide moral justification for punishing wrongdoers.

In this thesis, I will contend that such a theory exists: namely, a *contractarian* theory of punishment. I will argue that the moral justification for punishment arises from the hypothetical agreement of those to whom it applies. I begin in this chapter by discussing and endorsing Boonin’s definition of punishment, before then going on to outline his arguments against a number of existing theories of punishment. In Chapter 2 I provide a sketch of the general contractarian theory, and then suggest how it can be applied to the problem of punishment. I begin by noting that punishment is a means by which individuals can ‘purchase’ deterrence. Thus, if someone values the non-commission of an offence (say, because it threatens her person

---

1 This is analytically true on the definition of punishment given by David Boonin—a definition that I will endorse and adopt later in this chapter.
and property) then she can deter its commission by agreeing to some kind of punishment for that offence. In this way, individuals can benefit from certain punishment practices. I develop this theory in Chapter 3, arguing that where an individual can expect to benefit from a particular punishment practice, she will have good reason to agree to its implementation. I complete the theory in Chapter 4 by claiming that if an individual has good reason to agree to a punishment practice, she can be taken as giving her hypothetical agreement to its implementation. I contend that it is this hypothetical agreement that forms the moral justification of punishment. In Chapter 5 I consolidate and road test my theory, considering what kinds of punishment practices it would endorse for several typical crimes.

In Chapter 6 I go on to consider—and reply to—several potential objections to a contractarian theory of punishment. Specifically, I consider the problem cases of (1) the Value-Outlier, (2) the Vigilante, and (3) the Foole. Finally, in Chapter 7 I consider how the contractarian theory fares against the criticisms that Boonin levels against other theories of punishment. I show that my theory is uniquely placed to provide solutions to these objections. I contend that, for this reason, the contractarian theory of punishment provides an encouraging solution to the problem of punishment.

**What is Punishment?**

We cannot attempt to establish the moral permissibility of punishment without first providing an account of what punishment actually is. We may have some broad notion of the concept, but the following discussion will require a clear and rigorous definition of when something does—and

---

3 By ‘punishment practice”, I mean a particular punishment for a particular crime—e.g., a $50.00 fine for exceeding the speed limit by 5mph or less.
does not—count as an instance of punishment. Boonin does this by testing various conditions against what he takes to be “clear, paradigmatic instances of legal punishment” such as fines, imprisonment, bodily suffering, and death. The most obvious quality that these practices have in common is that they are each (in some way) bad for the person on whom they are inflicted. For this reason, Boonin claims that a natural starting point in generating a definition of punishment is to say that punishment *harms* the person who is punished—where harming someone means “making her worse off in some way, which includes inflicting something bad on her or depriving her of something good.” Importantly, this does not mean that an offender can receive no benefit from being punished. While the harm requirement holds that a certain treatment must harm an individual in order to count as punishment, it is “neutral on the further question of whether or not being subject to such harm might produce beneficial consequences in the future, including beneficial consequences that are great enough to outweigh (and perhaps even to justify) the immediate harmful ones.”

---

4 Boonin, p. 6.
5 This claim may, in fact, come to depend on our particular theory of welfare combined with certain facts about the individual—more on this shortly.
6 One concern with this harm requirement is how it might accommodate the masochist—such as the thief who is flogged for his crime, but for whom being beaten is an enjoyable experience. This might appear to be a case of a punishment that does not cause harm, thus showing that harm is not a necessary requirement for punishment. However, Boonin believes that we should simply claim that such cases of non-harm are not actually cases of punishment at all. He notes that there is a conceptual symmetry between punishment and rewards, and that we do not generally see a person as having received a reward if the treatment she receives does not end up benefitting her in some way. Since an individual cannot be rewarded without receiving a benefit—and since reward and punishment are symmetrical—this gives us good reason to hold that an individual can also not be punished without being harmed. But consider again the flogged offender who enjoys being beaten. If the above is true, then we cannot say that this offender has actually been punished. Does this then imply the controversial claim that flogging is not actually a punishment? Not necessarily. As Boonin notes, “there is a crucial difference between saying that a particular person has been subjected to a form of treatment that is a form of punishment and saying that this person has, in fact, been punished” (Boonin, p. 10). In order to better illustrate this response, Boonin draws an analogy with a sedative being administered to a patient who happens to be stimulated by that particular drug. The patient fails to be sedated, but the drug remains a sedative given certain general properties that it possesses (specifically, that it makes most people sleepy).
7 Boonin, p. 7.
8 Boonin, p. 7.
But while harm is a necessary condition for punishment, it is not sufficient. Further requirements must be met—the first of which is that the harm caused must be intentional. Boonin illustrates this by way of an example, comparing the case of an individual who is charged a fee for processing his marriage license, and another individual who is charged a fine for violating antipolygamy laws.\(^9\) The harm in both cases is of the very same kind—that is, a financial cost. Nevertheless, it is only the second kind of treatment that would likely be seen as ‘punishment’. The reason for this, claims Boonin, is that only in the second case was the harm intentional:\(^{10}\)

\[\text{When the state charges a fee for processing a marriage license, it understands that the cost imposes a harm on those getting married, but this is not its intention. Its intention is merely to recover the costs involved in processing the relevant paperwork, and it would charge the same fee even if, for some reason, couples getting married benefitted from paying it. When the state punishes someone, on the other hand, it inflicts various harmful treatments on him in order to harm him. It is not merely that in sentencing a prisoner to hard labour, for example, we foresee that he will suffer. Rather, a prisoner who is sentenced to hard labour is sentenced to hard labour so that he will suffer.}\]

It is this observation that leads Boonin to claim that punishment necessarily involves intentional harm. But this is still not sufficient. If I kick a random passer-by, I intentionally harm him—but I do not punish him. Punishment requires something more. Specifically, it requires intentionally harming someone because he previously committed a prohibited act.\(^{11}\) In the case of legal punishment, this can be further specified to “intentionally harming someone because he

---

\(^9\) Boonin, p. 12.  
\(^{10}\) Boonin, p. 13 (emphasis added).  
\(^{11}\) Boonin, p. 17.
previously did a legally prohibited act, which means that he is responsible for having done the act and that he had no valid legal excuse for doing so.”¹² Call this the retributive requirement. One concern with this condition is what we might say of those who are wrongfully convicted. Specifically, the retributive requirement seems to imply that those who are serving sentences for crimes they did not commit are not in fact being punished. Boonin’s response is to endorse this implication—while wrongfully convicted individuals are intentionally harmed, they are not in fact punished.¹³ This reasoning can be illustrated by considering the symmetry between reward and punishment. Suppose I give you ten dollars on the mistaken assumption that you have returned my wallet (when, in fact, you were the individual who stole it in the first place). In this case I clearly provide you with an intentional benefit, but this would not amount to a reward. I might believe that I am rewarding you, and I might intend the money to be a reward, but it is not a reward for the mere reason that you have done nothing to be rewarded for. If this is true (and, assuming that punishment and reward are symmetrical) an intentional harm cannot qualify as punishment unless the recipient has done something deserving of punishment.

Is retributive intentional harm sufficient for punishment? Not quite, argues Boonin. There are certain cases that will meet these three requirements, but that should not, it seems, count as cases of punishment. One such example might be where a criminal is branded for having committed a particular crime as part of a gang initiation.¹⁴ Such an action is a case of intentional harm carried out because an individual has committed a legally prohibited act. It does not, however, seem to count as a case of punishment. In light of this, further stipulation is needed. Boonin’s suggestion is to add a reprobative requirement—that is, a condition that to count as a

¹² Boonin, p. 17 (emphasis added).
¹³ Boonin, p. 18.
¹⁴ Boonin, p. 23.
punishment for an offence, the act must express official disapproval of the offender. The branding of the gang member does not do this. In fact, if anything, it expresses approval. Punishment carried out by the state however, does carry this element of reprobation.

Finally, Boonin notes that since the current debate focusses on the moral permissibility of legal punishment, a further condition—the authorisation requirement—should be added to our list of necessary and sufficient conditions. Specifically, something should only count as a case of punishment for the purposes of this discussion where it is “carried out by an authorised agent of the state acting in his or her official capacity.”

To summarise, Boonin claims that P’s act A is a legal punishment of Q for offence O if and only if:

1. P is a legally authorised official acting in his or her official capacity; and
2. P does A because P correctly believes that Q has committed O; and
3. P does A with the intent of harming Q; and
4. P’s doing A does in fact harm Q; and
5. P’s doing of A expresses official disapproval of Q for having committed O.

Thus, an act of punishment is—for Boonin—an act of authorised reprobative retributive intentional harm. It is practices of this specific kind for which Boonin believes traditional theories of punishment are incapable of providing a moral justification. As my argument is

---

15 Boonin, p. 22.
16 Boonin, p. 24.
17 My argument says nothing about the justification for alternative responses to wrongdoing such as rehabilitation. I would argue that a system of rehabilitation—if morally permissible—would operate parallel to a system of punishment, so that some cases of wrongdoing would merit a response from both systems, while other cases of wrongdoing may merit a response from only one (or possibly neither).
intended as a direct reply to Boonin, I will endorse and adopt this definition for the remainder of my thesis.

The Problem of Punishment

In the preface to his *The Problem of Punishment*, Boonin lays his cards on the table, stating that:18

Most people believe that if it is just and reasonable for the state to prohibit a given form of behaviour, then it is morally permissible for the state to punish those who persist in engaging in it. I don’t believe this. I don’t believe that it is morally permissible for the state to punish people for breaking the law.

Boonin’s position arises from the fact that punishment necessarily involves the state treating some of its citizens in ways that it would typically be wrong to treat others.19 Claire Finkelstein reinforces this observation, noting that “the institution of punishment involves acts that are normally highly morally objectionable,”20 and that because it includes “the deprivation of person liberty and the infliction of physical hardship,” punishment is “presumptively impermissible.”21

---

18 Boonin, p. ix.
19 Boonin, p. 213.
The problem of punishment lies in explaining how we can morally justify the intentional harm inflicted in the name of punishment. Specifically, a complete answer to the problem of punishment must explain:\(^{22}\)

1. Why the difference between those that punishment targets and those that it does not is morally relevant;
2. Why the difference between offender and non-offender is important enough to justify acts that will harm the offender; and
3. Why the difference between offender and non-offender is important enough to justify intentionally harming offenders.

According to Boonin, no current theory of punishment is able to provide such a set of explanations. In order to defend this claim, Boonin targets several versions of both the consequentialist theory of punishment (the view that punishment is justified insofar as its infliction would help deter the commission of future offences)\(^{23}\) and the retributivist theory of punishment (the view that punishment is justified when—and to the extent that—an offender deserves to suffer for the harm he has inflicted).\(^{24}\) He argues that each of these theories is found wanting, and that—for this reason—no moral justification has yet been given for the institution of punishment.

My intention in this thesis is not to re-litigate Boonin’s case against traditional theories of punishment, but to instead focus on the following question:

\(^{22}\)Boonin, p. 28.
If Boonin’s argument succeeds (meaning that consequentialism and retributivism both fail as theories of punishment) is there another theory that survives Boonin’s objections, and provides us with an alternative moral justification for punishment?

My strategy will be to assume that Boonin is correct in claiming that both consequentialism and retributivism fail as theories of punishment, but to then show that this does not give us reason to abandon punishment altogether—namely, because there is an attractive alternative to both consequentialism and retributivism.

Nevertheless, it is necessary to briefly examine the objections that Boonin does raise against traditional theories of punishment. Our purpose is to catalogue these complaints so that we might be wary as to whether any of these same arguments also apply to (and therefore discredit) the alternative theory I propose in the remainder of this thesis.

**Concerns with the Consequentialist Solution to the Problem of Punishment**

The first solution to the problem of punishment is provided by consequentialism. On this view, an act of punishment is justified because of its presumed good consequences. Consequentialist theories of punishment tend to focus largely on deterrence—that is, on punishment’s ability to discourage the future commission of crime, and therefore the creation of harm. Boonin addresses two different versions of the consequentialist approach—act utilitarianism, and rule utilitarianism—and outlines the difficulties with each.

25 Boonin, p. 37.
26 Though some consequentialist theories may take other future benefits—such as the rehabilitation of the offender—into account.
Act Utilitarianism

The act-utilitarian solution to the problem of punishment holds that punishing people for breaking the law is justified because “it best promotes (or can reasonably be expected to best promote) human happiness as well as well-being.” The punishment of offenders is seen as being useful—of having a utility. Specifically, this utility is the deterrence of future crimes. By fining an individual for speeding, we deter others from speeding, and therefore reduce the sum total of harm caused by drivers who go too fast. Herein lies the moral justification for punishment, according to the act utilitarian.

While intuitively plausible, there are a number of serious problems with this approach to punishment. The most commonly raised of these is the claim that act utilitarianism may, on occasion, justify punishing an individual who has not actually committed a crime. Call this the Punishing the Innocent Objection. Suppose a vicious murder has been committed in a small town, and that the police have resigned themselves to the fact that they will be unable to discover the identity of the offender. The citizens of the town are not content with this however, and plan to riot. During these riots, mass destruction of property will occur, along with the potential lynching of several suspicious—yet ultimately innocent—individuals. In such a case, it would seem that act utilitarianism would endorse framing and punishing a single innocent individual in order to avoid the much wider harm that will occur if no individual is punished at all.

---

27 Boonin, p. 39.
28 Boonin, p. 41.
29 The utilitarian might make an initial response by noting that—on our definition of punishment—it is impossible (for the same reasons discussed earlier) to punish the innocent. While technically true, this does nothing to address the central thrust of the objection. As Boonin notes, the objection could merely be modified terminologically to state that the justification that the act-utilitarian solution provides “for fining, imprisoning, and executing guilty people as forms of punishment would equally justify fining, imprisoning, and executing some innocent people as forms of ‘pseudopunishment’ or ‘quasi-punishment.’” (Boonin, p. 44).
Boonin notes that there are two general responses that an act utilitarian might give to this objection. The first is to simply accept the implication—agreeing that it is morally permissible to punish innocent individuals in certain suitable circumstances. For many, however, this is an unacceptable position. The only alternative, then, is to deny the very implication that act-utilitarianism would endorse the occasional punishment of innocent individuals. This would be done by arguing that the cost of punishing the innocent would be greater than the rendered benefits. The chief way of doing this is by appealing to the necessary costs in keeping the practice a secret, because if the general public discovered that an innocent person—not the actual perpetrator—had been punished it would “break down all trust and respect for the law,” and “the evil of such a breakdown would outweigh any good that might be obtained by punishing the merely supposedly guilty” (or innocent).\textsuperscript{30}

Does this response succeed? Boonin thinks not. For one, this claim regarding the cost of secrecy is highly contingent on the story being told. It may very well be true in some scenarios—and, in those cases, act-utilitarianism will not provide a justification for the punishment of an innocent. There are many cases, however, where the cost of secrecy is not so high (at least, not high enough to outweigh the benefits to be gained from the practice). This may be due to either the (1) chances of discovery being minimal, or (2) the potential harm caused by discovery being particularly low.\textsuperscript{31} Further, not all cases of punishing the innocent even require secrecy. Suppose, for example, that a parent repeatedly committed a crime (such as stealing food) for the benefit of her child. Punishing the parent would likely not be a highly effective deterrence (a parent who is willing to break the law for the benefit of her child is unlikely to be deterred by the prospect of punishment). What might be a highly effective deterrence is to punish the parent vicariously

\textsuperscript{31} Boonin, p. 45.
through her child. This would amount to the punishment of an innocent which would require no secrecy (in fact, the very opposite would be preferred—wide knowledge of the punishment would serve to enhance its deterrent effect). Act utilitarianism, it seems, would endorse such treatment.

Just as act utilitarianism may endorse the punishment of innocent individuals; it might also require us to fail to punish those who are guilty. Call this the **Not Punishing the Guilty Objection**.\(^{32}\) Perhaps the offender is a pillar of the community, or responsible for the provision of some kind of unique and incredibly valuable service without which the community will suffer immensely. So long as the harm caused by punishing the offender exceeds the benefit gained in deterrence, act utilitarianism will implore us to refrain from punishing the offender.

Act utilitarianism might also prescribe punishments that seem highly disproportionate. Call this the **Disproportionate Punishment Objection**.\(^{33}\) A petty crime such as stealing candy may be very difficult to deter, and therefore necessitate an incredibly strong punishment. Likewise, it might be discovered that a serious offence is carried out merely for the purposes of cheap entertainment (Boonin uses the example of arson),\(^{34}\) and that a relatively small fine might be all that is needed to deter the offence when, in fact, our moral intuitions seem to endorse a much heftier penalty.

Two further problems arise for an act utilitarian theory of punishment. For one, this approach does not bestow any sort of relevance on mitigating or aggravating factors. Call this the **No Excuses Objection**.\(^{35}\) Legal systems commonly recognise that certain facts about the circumstances and mental state of the offender should have some bearing on the severity of that

---

\(^{32}\) See Boonin, pp. 52-54 for more.

\(^{33}\) See Boonin, pp. 54-58 for more.

\(^{34}\) Boonin, p. 57.

\(^{35}\) See Boonin, pp. 58-60 for more.
individual’s punishment. Under the act-utilitarian system however, these factors play no role. The only concern is the extent to which the offender’s punishment can be used to deter future commissions of the offence in question.

This utilisation of the offender gives rise to the further concern of using people to achieve some end. Call this the **Treating People as a Means Objection**. Boonin notes that the act-utilitarian might respond to this particular objection by claiming that while it is impermissible to use an *innocent person* as a means to an end, it is permissible to use an *offender* in this way. However, while such a claim may be true, it will require a justification that is non-utilitarian in nature.

It might turn out to be true, for example, because guilty people deserve to be treated this way, or have consented to be treated this way, or have forfeited their right not to be treated this way. But if a defender of punishment has to appeal to such considerations, he is at that point abandoning a utilitarian solution in favour of a nonconsequentialist one.

Alternatively, an act-utilitarian might claim that the state routinely treats people as means, with little objection. An individual suffering from a virulent and easily communicable disease might, for example, be forced to undergo serious restrictions on her freedom of movement in the interests of public health. But Boonin notes that these cases typically miss the distinction between *intending* harm, and merely *foreseeing* harm. The harm caused to a patient by her quarantine is incidental—and would no doubt be avoided if at all possible. It is a regrettable by-

---

36 See Boonin, pp. 60-62 for more.
37 Boonin, p. 61.
38 Boonin, p. 61.
39 Boonin, p. 61.
product of the end of protecting others from the disease. Contrast this with the case of punishment, where the harm caused is (by definition) intentional.40

Evictions, quarantines, and military conscription do not involve intentionally harming some people in order to benefit others. Rather, they involve intentionally doing acts that foreseeably cause some harm to some people and provide greater benefits to many others. It is perfectly consistent for the critic of the act-utilitarian solution to accept all of these practices as morally permissible while still insisting that the act-utilitarian solution impermissibly treats offenders as mere means to promoting the public good.

Given these allegedly insurmountable concerns, argues Boonin, act-utilitarianism must be abandoned as a viable theory of punishment.

**Rule Utilitarianism**

The utilitarian should not be counted out yet, however. One move they might make is to switch from act-utilitarianism to *rule*-utilitarianism. Rule-utilitarianism importantly identifies the difference between “justifying a practice as a system of *rules* to be applied and enforced, and justifying a particular *action* which falls under these rules.”41 On this view, a particular rule of punishment will be justified because following that *rule* best promotes (or can reasonably be expected to promote) human happiness or well-being.

---

40 Boonin, p. 62.
The appeal of the rule-utilitarian approach is in its alleged ability to avoid certain counter-examples—specifically those relating to the punishment of the innocent. In order to illustrate this, John Rawls discusses a practice of “telishment”—an institution which is such:

[T]hat the officials set up by it have authority to arrange a trial for the condemnation of an innocent man whenever they are of the opinion that doing so would be in the best interests of society. The discretion of officials is limited, however, by the rule that they may not condemn an innocent man to undergo such an ordeal unless there is, at the time, a wave of offences similar to that with which they charge him and telish him for.

Rawls then argues that when comparing the institution of punishment with the institution of punishment + telishment, utilitarianism will almost certainly favour the former:

Once one realises that one is involved in setting up an institution, one sees that the hazards [of telishment] are very great. For example, what check is there on the officials? How is one to tell whether or not their actions are authorised? How is one to limit the risks involved in allowing such systematic deception? How is one to avoid giving anything short of complete discretion to the authorities to telish anyone they like? In addition to these considerations, it is obvious that people will come to have a very different attitude towards their penal system when telishment is adjoined to it. They will be uncertain as to whether a convicted man has been punished or telished. They will wonder whether or not they should feel sorry for him. They will wonder whether the same fate won’t at any time fall on them. If one pictures how such an institution would actually work, and the enormous risks involved in it, it seems clear that it would serve no useful purpose. A utilitarian justification for this institution is most unlikely.

42 Rawls (1955), p. 11.
The rule-utilitarian’s response to the Punishing the Innocent Objection is therefore as follows: to accept the implication that punishing an innocent individual will, on occasion, maximise utility—but to respond by claiming that an *institution* which allowed such actions would, on the whole, reap a lesser utility than an institution which did not.

Is this enough for the rule-utilitarian to escape the objection? We might say “no” by objecting to the optimism with which Rawls describes the institution of telishment. Perhaps the risks are not so great, nor the opportunities for abuse so endemic. But let us be charitable, and assume that things are exactly as Rawls has described—that is, that a system of punishment + telishment will, on a utilitarian assessment, fail to be justified over a system of mere punishment. Even if this is the case, what does the existence of a justified rule mean for the justification of a lone action? Rules are—as the old adage insists—made to be broken. In other words, we must ask what reason judges have to *always* follow the rules that the legislature sets down. As Boonin notes:

> The mere fact that adopting a rule at one point in time might be the best thing to do from a utilitarian point of view at that point in time does not entail that following that rule at every subsequent point in time will be the best thing to do from a utilitarian point of view at those points in time. And if one now finds that one can produce more overall good by breaking a rule that it made the most sense on utilitarian grounds to adopt at an earlier point in time, then it should seem clear that a utilitarian must urge you now to break the rule even though accepting the rule was initially justified on good utilitarian grounds.

---

44 For more on this—what Boonin refers to as the “Rule Content Problem”—see Boonin, pp. 70-72.
45 Boonin, p. 70 (emphasis added).
If this is so, then the utilitarian merely comes full-circle. Her normative views may endorse an institution of punishment over a system of punishment + telishment. But when certain ideal circumstances arise—when the mob is baying for blood, and the only way to avoid serious and widespread harm to a massive number of people is to sacrifice a single innocent individual—then utilitarianism will scream out for such action to be taken.

Non-Utilitarian Consequentialism

There are, of course, varieties of consequentialism beyond utilitarianism. Instead of systematically addressing each, however, Boonin notes that the central problem of utilitarianism (that is, that which makes it eternally vulnerable to objections such as Punishing the Innocent and Not Punishing the Guilty) is in fact systemic to all consequentialist theories. Stated plainly, the problem arises from the fact that all consequentialist theories are—analytically—forward-looking in nature, and that on any account of the good which is aimed at the future: ⁴⁶

[T]here will be cases in which more future good will be produced by punishing some innocent people and not punishing some guilty people, in which more good will be produced by punishing some guilty people too much and others too little, and in which people will be harmed merely to produce a greater overall amount of good.

For this reason, argues Boonin, we must look elsewhere for a satisfying solution to the problem of punishment.

⁴⁶ Boonin, p. 84.
Concerns with the Retributivist Solution to the Problem of Punishment

While the consequentialist solution to the problem of punishment is essentially forward-looking, the alternative—retributivism—is essentially backward-looking. Stated plainly, retributivism holds that “committing an offence in the past is sufficient to justify punishment now, whether or not this will produce any beneficial consequences in the future.” There are a variety of foundations upon which a retributivist theory of punishment might be built. Boonin discusses the three most promising.

Desert-Based Retributivism

Desert-based retributivism holds that punishing people for breaking the law is morally permissible because such people deserve to be punished. Support for this theory is often garnered by arguing from cases: When we consider instances of extreme evil, for example, most of us will agree that the state of affairs in which an offender is punished for his offence is better than the state of affairs in which he is not punished. Something about the offender’s past behaviour makes it appropriate for us to now respond by visiting certain consequences upon him.

Desert-based retributivism is not without its problems, however. For one, it appears to fall prey to the familiar Punishing the Innocent Objection. Some individuals may do many things deserving of punishment without actually breaking the law. Consider Boonin’s example of the racist.

---

47 Boonin, p. 85.
48 Boonin, p. 85.
49 Boonin, p. 87.
50 Boonin, p. 99.
[She] does everything she is legally permitted to do to insult black people. If she is legally allowed to play racist songs, she does. If she is allowed to throw a party celebrating the fact that hundreds of black people died in an earthquake, she does. She refrains from burning crosses on people’s lawns if she must, and she avoids lynching black people only because she is afraid of being caught.

This individual is legally innocent. Yet, on many common standards of desert, she will deserve to suffer for her actions. And if desert-based retributivism is true, this deservingness will provide a moral impetus for us to punish her—thereby endorsing the punishment of the (legally) innocent.

This problem runs in the opposite direction, too, with desert-based retributivism apparently also suffering from the Not Punishing the Guilty Objection. The individual who breaks an emissions law by driving her smoke-belching car to the hospital to save her friend’s life is legally guilty—but few will believe she has acted so as to deserve punishment.\(^{51}\) And if desert-based retributivism is true, this absence of deservingness will mean that there is no moral justification for punishing her. A similar scenario might occur where an individual innocently and non-negligently breaks a law (say by accidentally entering the wrong home believing it to be his own)—making him legally guilty, but (on desert-based retributivism) undeserving of punishment once again.\(^{52}\)

Now these objections may not necessarily appear insurmountable for desert-based retributivism. We may in fact agree with the answers that desert-based retributivism provides us in the scenarios outlined above. Perhaps we do believe that the vehement racist should suffer (despite her innocence), and that the life-saving gas-guzzler should not (despite her guilt). There are, in fact many cases in which we do not expect legal and moral desert to coincide fully. Take

\(^{51}\) Boonin, p. 94.

\(^{52}\) Boonin, p. 96.
the example of adultery. Many Western cultures see this behaviour as morally impermissible, yet few go so far as to make it deserving of legal punishment. But this observation is the very thing which marks the death knell of desert-based retributivism as a solution to the problem of punishment. Legal guilt and legal punishment must coincide. We cannot have an institution that punishing some crimes and not others, and punishes some actions that are not crimes at all. The fact that an individual *morally deserves to suffer* will not be an accurate test for when she should *legally be punished*.

There is also a second, deeper, problem with desert-based retributivism. As noted above, the very foundation of this solution is the observation that the state of affairs in which an offender gets what she deserves is intrinsically better than the state of affairs in which she does not. In order for this approach to get off the ground, however, it must be shown that this intrinsic value of a certain state of affairs necessarily entails the claim *that it is morally permissible to impose that state of affairs on a person*. This is far from obvious, however. As Boonin notes:\(^53\)

\[\text{The state of affairs in which only one innocent person dies is much better than the state of affairs in which five innocent people die… but this does not in itself establish that it is morally permissible to kill one innocent person to prevent five others from dying.}\]

Thus, the desert-based retributivist needs to provide more justification for punishment than to simply appeal to the intrinsic value of the offender being held accountable for her actions.

\(^{53}\) Boonin, p. 102.
Forfeiture-Based Retributivism

Alternatively, retributivists might instead ground their solution in a more general theory of rights. This approach—called forfeiture-based retributivism—begins with the following claim:

If $P$ violates $Q$’s right to $X$, then $P$ forfeits $P$’s own right to $X$ (or perhaps instead forfeits some equivalent right or set of rights).

The forfeiture-based retributivist must then provide grounds for such a claim. Ordinarily, this is done by pointing to an important relation between moral rights and moral duties. The affirmation of moral rights seems—on most accounts—to entail the affirmation of moral duties (affirming one’s right to life, for example, seems to entail affirming a duty to respect the right to life of others). If this is the case, then it follows that negating one’s moral duties might also involve negating one’s moral rights.

But this appeal to a connection between rights and duties is problematic. Very young children and infants, foetuses, nonhuman animals, and people with severe mental disorders do not have moral duties, yet many of us would hold that (at least some of) these beings should be rightfully afforded moral rights. One potential response by the rights-duties defender might be to say that the duty not to violate a right simply amounts to a duty to refrain from violating that right when one is able to do so. This, however, makes the rights-duties connection meaningless, as everything without rights will have duties as well. As Boonin notes, “we would

---

54 Boonin, p. 106.
56 Boonin, p. 107.
have to say… that a grain of sand has a duty not to break its promises, since if it had the ability to make and break promises, it would be wrong to break them.”58

A number of further objections plague the forfeiture-based retributivist solution. For one, the consideration of a number of cases soon shows that the “right-for-a-right” approach leads to some deeply troubling conclusions: the torturer forfeits his right not to be tortured, the rapist forfeits his right not to be raped, the government official who illegally prevents members of a religious denomination from congregating forfeits his own right to religious freedom, and the judge who receives a bribe forfeits his right to receive a fair sentence.59 Some forfeiture-based retributivists might bite the bullet, saying that the theory provides the correct answer in these cases. More likely, however, is that proponents of the solution will instead appeal to some notion of “equivalent” rights. An offender who violates a right of her victim will therefore stand to lose either that same right, or some equivalent right / set of rights. The torturer, for example, might not lose his right to not be tortured, but may instead lose his right to freedom of movement for a time (i.e., via incarceration). Unfortunately, this response is also not without its problems. If removing an individual’s right to not be tortured seems repugnant, but removing his right to freedom of movement is genuinely equivalent, then we should be equally outraged by the latter punishment as the first. As Boonin puts it:60

[I]f a given right really is equivalent to another, then if it is unacceptable to deprive someone of one right, it must be equally unacceptable to deprive him of the other. If, on the other hand, we have no qualms about depriving someone of a certain right but would strongly resist depriving him of some other right, then this fact in itself demonstrates that the rights are not equivalent.

---

58 Boonin, p. 108.
59 Boonin, p. 110.
60 Boonin, p. 111.
An additional concern with forfeiture-based retributivism stems from the theory’s inability to dictate the period for which an offender loses his or her right/s. In many cases, a permanent loss of a right will appear extreme—an individual who, say, falsely imprisons his victim for fifteen seconds should not (at least on most reasonably views of punishment) suffer a life-sentence of imprisonment. Yet forfeiture-based retributivism fails to give us any clear indication of where (or, more importantly, how) a line is to be drawn with regards to the temporal extent of a punishment.

Third, a blanket removal of an offender’s right seems to license far more than just the institutional response of the state. Specifically, it opens the doorway for private retaliation. If forfeiture-based retribution is true, then a murderer loses his right to life. However, the loss of this right seems to permit any individual to take the life of the offender without suffering punishment (as the private retaliator will not be violating the offender’s right to life, since the offender had no such right).

Forfeiture-based retributivism even sees the return of the Punishing the Innocent Objection. In order to explain the moral (not merely legal) permissibility of punishment, the forfeiture-based retributivist solution must focus on moral—not legal—rights. But there are many moral rights that are either (i) not enshrined in law, or (ii) enshrined in law to only a limited extent. Boonin provides examples such as being lied to by a partner (violating an individual’s right not to be lied to), or having fries stolen from your plate by a friend (violating your property rights). Neither such case involves legal guilt, but both—if forfeiture-based retributivism is true—provide grounds for punishment.

---

61 For more, see Boonin, pp. 112-114.
62 Boonin, p. 113.
63 For more, see Boonin, pp. 114-115.
64 For more, see Boonin, pp. 115-117.
Finally, and perhaps most importantly, there is a concern about whether an institution built on forfeiture-based retributivism would even amount to punishment. The reason for this is as follows: We do not intentionally harm an individual by taking something from a person that she was not entitled to in the first place. I do not intentionally harm my neighbour by taking back the newspaper she ‘accidentally’ mistook as her own and removed from my front doorstep. She is harmed (she is now without a newspaper) but this harm was not intentional. I foresaw her harm as a by-product of my remedying a wrong, but I did not intend it. So too is it with punishment. If an offender forfeits a right to something (by violating those same rights for another) then she is not harmed when the state acts so as to remove the very thing she no longer has a right to. But, since intentional harm is a necessary condition for punishment, the removing of something to which someone has no right cannot amount to punishment.

Consider an example provided by Boonin to help clarify. Suppose that you renege on a contract to make monthly rental payments for a car. Suppose, further, that such a failure of payment is clearly stipulated as a forfeiture of all rights to the vehicle. The dealership takes you to court, and the judge forces you to give up the car. In doing so, the judge:

…is doing something that harms you, and she recognises that this harms you. But—and this is the crucial point—the judge is not taking the car away from you with the intention of causing you harm and making you suffer. She is merely taking the car away from you in order to make sure that you do not keep something you have no right to possess. She foresees that in enforcing the terms of the contract you will be harmed, but harming you is not her intention. And because of this important fact, what the judge does to you when she takes the car cannot be classified as punishment.

---

65 Boonin, p. 118.
If this is true, then the same would apply to every other case of purported punishment handed down by the state. In every case, a judge would merely be removing from an individual what she has already lost the right to possess. And such removal will never—for the reasons just given—amount to intentional harm, and therefore not count as punishment.

*Fairness-Based Retributivism*

One final bastion for retributivists might be to instead ground their solution in a general principle of fairness. This theory—called “fairness-based retributivism”—begins its solution by framing the problem of punishment in terms of distributive justice:

66 Boonin, p. 120.

Punishment… involves imposing certain burdens on some people that are not imposed on others.

The problem of punishment, on this account, involves explaining how it can be fair to distribute these burdens so inequitably.

Once couched in this way, the approach of fairness-based retributivism becomes clear. Specifically, it will see offenders as:

67 Boonin, p. 120.

…enjoying an unfair distribution of benefits as a result of having committed an offence, so that imposing on them a punitive harm will restore the overall distribution of benefits and burdens to its previous and presumptively fair level.
By ‘benefit’, we do not mean the potential fruits of the offender’s crime—though she may indeed benefit in certain ways. Suppose an offender steals several thousand dollars of home theatre equipment from a department store. This individual now clearly enjoys a benefit that other law-abiding people do not—namely, the acquisition of expensive material goods for free. The court would, subsequently, at least require the return of these goods to the department store. But for the reasons mentioned in the above discussion of forfeiture-based retributivism, this would not amount to punishment (as no intentional harm is done by forcing an individual to give up what they already have no right to). The benefit which the home-theatre-thief unfairly enjoys is instead her non-acceptance of the burden of self-restraint assumed by those who obey the law.\textsuperscript{68} In breaking the law, the offender “has enjoyed a freedom from such restraint that others have not.”\textsuperscript{69} As a result, fairness-based retributivism holds it to be morally justifiable for the state to punish her such that she receives a burden sufficient to nullify her ill-gained benefit. When an offender is—for example—put in prison, the liberty that she is now deprived of is intended to compensate for the extra amount of liberty that she unfairly took for herself in committing the offence.

While such an approach may make sense in the case of theft, it is more difficult to understand how it could apply to serious offenses such as murder and rape. For one, the general strategy of fairness-based retributivism is to claim that punishment is justified for a crime, because offending amounts to enjoying benefits that others have refrained from—a sort of “free-riding.” But to claim that punishment for rape and murder is justified merely on the basis that these offences count as free-riding is—in the words of R. A. Duff—“perverse.”\textsuperscript{70} Additionally, there is the further concern of how we can even make the offences reach the standard of free-

\textsuperscript{68} Boonin, p. 121.
\textsuperscript{69} Boonin, p. 121.
riding. The general fairness-based retributivist claim is that offenders enjoy the benefit of not exercising the self-restraint exhibited by law-abiding citizens. But this seems incorrect in the case of serious offences. Most people have no desire to murder or rape, so adhering to the law requires no self-restraint whatsoever.\footnote{Boonin, p. 123.}

One response that might be given by the fairness-based retributivist is to claim that offenders are not free riders because the rest of us incur a serious cost in adhering to a specific law (like those against murder or rape), but because we incur a serious cost \textit{in adhering to the law in general}.
\footnote{Boonin, p. 124.} According to this response, non-offenders suffer a detriment by adhering to a wider system of laws, including refraining from doing \textit{some} things which would benefit them (for example, getting to work faster by speeding). But this response simply invites another problem. If the standard that an offender fails to meet is merely “the obligation to adhere to the law in general” then all law-breakers are free-riders to the same extent. If this is true however—and if the moral permissibility of punishment derives from the incidence, and extent of, a case of free-riding—then all law-breakers \textit{will be punishable to the same extent}.
\footnote{Boonin, p. 124.} Unless we are willing to punish the petty thief in the same way as the serial killer, such an entailment is unacceptable.

One way around this might be to adopt the strategy of George Sher, who claims that the true measure of an offender’s unfair extra burden lies in the seriousness of the moral prohibition that he has violated.
\footnote{George Sher, \textit{Desert} (Princeton, NJ: Princeton University Press, 1987), p. 81.} But there are several problems with such a response. First, there is no independent reason to accept the claim that one benefits more from violating morally greater prohibitions.
\footnote{Boonin, p. 127.} In fact, there are many scenarios that would seem to act as counterexamples to such a claim. An offender might benefit far more from the theft of ten thousand dollars than he
would from the murder of a homeless man—yet the latter offence clearly involves the violation of a much more serious moral prohibition. Second, Sher’s response only applies to violations of law that prohibit *morally objectionable* behaviour. It doesn’t cover, for example, driving without a license or building an addition to one’s house without a permit.\(^{76}\)

Another difficulty for fairness-based retributivism is how it will manage to deal with cases of previously victimised offenders. Victims are already suffering from an unfair distribution of goods—namely because they have less than everyone else. If they commit an appropriate offence, they may then bring themselves back up to their original level (and the level of all other citizens). In such a case, the offender does not possess an unfair (larger) distribution of goods than his neighbours, so it is unclear how fairness-based retributivism will justify punishing him.\(^{77}\)

Further additional problems remain for fairness-based retributivism: First, the No Excuses Objection rears its head again, with this theory of punishment unable to take into account mitigating circumstances.\(^{78}\) Second, the theory is yet to provide a reason to show that one person’s free riding on the sacrifices of others entails that it is morally permissible to coercively extract payment from her for doing so.\(^{79}\) Third, we encounter the The Punishing the Innocent Objection once more, as if it is morally permissible for the state to punish offenders because they are free-riders, then it is morally permissible for the state to punish *non-offenders* who are free-riders for the same reason (thus permitting the punishment of those who are not legally guilty).

\(^{76}\) Boonin, p. 129.
\(^{77}\) For more, see Boonin, pp. 135-137.
\(^{78}\) For more, see Boonin, pp. 137-138.
\(^{79}\) For more, see Boonin, pp. 138-139.
Finally, there is—as there was with forfeiture-based retributivism—a concern about whether or not what flows from the justification will even count as punishment. As noted earlier: to take something from an individual that she is not entitled to have in the first place does not count as intentional harm, and therefore cannot qualify as punishment. In the current scenario, if the offender is not entitled to his unfair distribution of goods, then removing these goods from him will not intentionally harm him, and therefore will not count as punishment under the definition outlined at the beginning of this chapter.\textsuperscript{80}

A New Solution to the Problem of Punishment

This catalogue of concerns with both consequentialist and retributivist theories of punishment leads Boonin to the conclusion that it is never morally permissible for the state to punish people for breaking the law. I wish to avoid this conclusion. As noted earlier, however, it is not my intention to relitigate Boonin’s case against the traditional approaches to punishment. It may very well be the case that one or more theories are able to escape his concern. This, however, is a job for the consequentialist or the retributivist.

I will employ a different strategy. I will grant Boonin’s claim that both approaches fail to provide a plausible account of the moral permissibility of punishment, but show that even in spite of this, we need not abandon punishment as a practice altogether. Instead, I will propose a new theory of punishment—one that is neither consequentialist nor retributivist, but instead contractarian in nature. I will show that this theory escapes the kinds of objections Boonin levels

\textsuperscript{80} For more, see Boonin, pp. 141-143.
at traditional theories of punishment, and therefore provides a viable alternative to his strong eliminativism concerning punishment.  

Boonin does briefly consider—and reply to—one specific version of a “consent solution” to the problem of punishment: that put forward by C.S. Nino in “A Consensual Theory of Punishment,” Philosophy & Public Affairs, Vol. 12, No. 4 (1983), pp. 289-306. However, Nino’s solution conceptually differs from my own theory in two fundamental ways: Firstly, Nino’s solution is based in tacit consent, while my theory will be based in hypothetical consent. Secondly—and more importantly—there is a fundamental difference between Nino’s theory and my own in what is being agreed to. Under Nino’s theory, the thing that is being agreed to is a specific instance of punishment for a specific crime—an agreement that is derived directly from the individual’s breaking of the law. Under my theory, however, the thing being agreed to is the more general punishment practice under which an individual could permissibly be punished if she goes on to commit a relevant crime. As I will explain in Chapter 4, this agreement will be based upon an individual’s expectation of benefit under the practice in question. This conceptual difference in what is being agreed to (and how that agreement is found) is sufficient to render most of Boonin’s objections to Nino’s theory inapplicable to the theory I will propose here. I will, however, turn to consider any residual concerns in Chapters 6 and 7.
Boonin argues that both consequentialism and retributivism fail as theories of punishment, and that—for this reason—*it will never be morally permissible to punish people for breaking the law*. This may seem an audacious claim. But, to be fair to Boonin, we must note that this simply means that it will never be morally permissible to carry out acts of *authorised reprobative retributive intentional harm*. This does not mean that we would have to refrain from responding to wrongdoing altogether. Certain responses—particularly those based on the grounds of rehabilitation or protection—would still be permissible. It may even still be appropriate to *harm* an offender, provided that the harm is not caused intentionally. Boonin’s specific strategy is to employ a system of victim restitution where, insofar as possible, an offender is required to restore his or her victim/s to the same status they would have occupied had the offence never been committed.\(^82\) Forcing an offender to pay restitution will, in most circumstances, *harm* the offender—but this restitution is not demanded with the intention of causing this harm, but rather with the intention compensating the victim. For this reason, it does not qualify as ‘punishment’ as defined in the preceding chapter.

I wish to propose an alternative response to Boonin’s argument. I do not believe that the failure of consequentialism and retributivism as theories of punishment necessitates the abandonment of punishment altogether. Instead, I will argue that there is another theory that escapes Boonin’s objections, and that provides us with an alternative moral justification for punishment—namely, a *contractarian* theory of punishment. Before going on to outline this

---

\(^82\) See Boonin, pp. 213-275.
view, however, it is important to first explain exactly what I mean by ‘contractarianism’—
including distinguishing it from its often-confused theoretical sibling of contractualism.

“Contractarianism”

For the purposes of this thesis, I use the term ‘contractarianism’ to refer to the specific theory
outlined by Thomas Hobbes in *Leviathan*, and later developed by David Gauthier in *Morals by
Agreement*. Contractarianism is just one member of a much larger family of theories that make
up the social contract tradition. Generally, these theories share the view that society is “a
cooperative venture for mutual advantage.” On this view, society is seen as resulting from
certain kinds of agreements between rational agents who recognise that they will each fare better
under the terms of social interaction than they would in its absence.

Within this society, morality is taken as being modelled on—or, in some cases, directly
derived from—“mutually agreeable reciprocity or cooperation between equals.” In this way,
the concepts of ‘right’ and ‘wrong’ are understood as “rules… that underlie the broadest possible
cooperation… [between] all competent human or rational agents.” Thus, whether an action is
morally permissible will depend on whether that act “accords with or violates principles that are,
or would be, the object of a suitable agreement between equals.” Divisions between social
contract theories occur when it comes to explaining how (and from where) these agreements
arise.

---

87 Stephen Darwall (ed.), *Contractarianism / Contractualism* (Malden, MA: Blackwell Publishing Company, 2003),
p. 1.
88 Darwall, p. 3.
89 Darwall, p. 1.
Hobbes begins the outline of his own (contractarian) social contract theory with a thought experiment. He invites the reader to imagine a world without morality or law—a world in which every man is governed solely by his own reason, the primary focus of which is the preservation of the individual.\footnote{Hobbes, Part I, Chapter 14, p. 189.} He claims that in such a condition, there would be nothing a man could use “that may not be a help unto him, in preserving his life against his enemyes”\footnote{Hobbes, Part I, Chapter 14, pp. 189-190.} and that for this reason, every man would have “a Right to every thing; even to one anothers body.”\footnote{Hobbes, Part I, Chapter 14, p. 190.} Hobbes claims that this natural state of man would be “a condition of Warre of every one against every one”\footnote{Hobbes, Part I, Chapter 14, p. 189.} in which life was “solitary, poore, nasty, brutish, and short”\footnote{Hobbes, Part I, Chapter 13, p. 186} and there could be “no security to any man… of living out the time, which Nature ordinarily alloweth men to live.”\footnote{Hobbes, Part I, Chapter 14, p. 190.}

Herein lies the paradox.\footnote{I use this term loosely—not to assert that there is some logical contradiction in our description of the natural state of man, but rather to draw attention to the way in which the relentless pursuit for one particular goal (namely, self-preservation) leads to a state of affairs that specifically works against the attainment of this very goal.} Although the natural condition results from every individual’s rational pursuit of his own goals, it leads to a state that is (as David Gauthier eloquently puts it) “inimical to preservation.”\footnote{Gauthier (1987), p. 159.} This thought experiment helps to identify an important feature of interactions between rational individuals—namely, that:

[while] each person’s pursuing his interests may result in their actually being best promoted, \textit{given} the conduct of others, it doesn’t follow that \textit{everyone’s} pursuing his own interests, rather than acting on some principle other than self-interest, will actually result in everyone’s (or even anyone’s) interests being best promoted.\footnote{Darwall, p. 2 (original emphasis).}
A scenario in which the collective pursuit of self-interest leads to an outcome that is worse for each person involved is an example of the collective action problem. A theoretical example of this problem can be found in the infamous case of the Prisoner’s Dilemma:

Two individuals are jailed on suspicion of robbery. The district attorney tells each that she lacks enough evidence to convict either of robbery, but can easily convict both of breaking and entering with a sentence of one year. She offers each a deal: if one confesses and his partner doesn’t, the confessor will go free and the partner will get twenty years. If both confess, both get five years.

Suppose that you and your friend find yourself in this situation—isolated in different jail cells, and contemplating your options. You have two options to choose from: either to confess to the robbery, or to remain silent. Depending on what your friend decides to do, there are four potential outcomes for you. First, you might walk free. Most likely, this is your first-ranked preference. Alternatively, you might find yourself imprisoned for either one, five, or twenty years—presumably representing your second-, third-, and fourth-ranked preferences in that order. What, then, should you do? Neither you, nor your accomplice, will have any idea of what the other decides to do. Your decisions are made entirely independent of each other. Given that observation, it seems that your best course of action is to act so that you maximise your outcome regardless of your friend’s decision.

In order to calculate this, you need to consider the two possible outcomes of your friend’s decision: she either confesses, or she does not. Suppose that your friend confesses. If you also confess, you will receive five years in prison (your third-ranked option). If you remain silent, you will receive twenty years in prison (your fourth-ranked option). Since five years is preferable to

---

99 Darwall, p. 2 (original emphasis).
100 Darwall, p. 2.
twenty years, it is clear that if your accomplice confesses, you are better off by confessing as well. Consider, now, what would happen if your accomplice instead chooses to remain silent. If you also remain silent, you will receive one year in prison (your second-ranked option). If, however, you decide to confess, you will walk free (your first-ranked option). Since walking free is preferable to one year in prison, it is clear that if your accomplice remains silent, you are still better off by confessing.

Thus, whatever your friend does, you will be better off by confessing. It seems, then, that you have good reason to confess. But then the realisation hits you that your friend is also a very smart individual, and that she has most likely calculated that *she has the very same reasons to confess*. This is where the dilemma arises. By both confessing, you each end up with five years in prison—the third-ranked option for each of you. If you had both simply remained silent, however, you would each walk free after serving only one year in prison. So while each of you confessing is “likeliest to achieve the agent’s best outcome when taken individually,” 101 taken together these actions “yield an outcome that is worse for each.” 102

Gauthier notes that there are many other situations like this in which “if each person chooses what, given the choices of the others, would maximise her expected utility, then the outcome will be mutually disadvantageous in comparison with some alternative.” 103 What, then, is the solution? Let us return momentarily to Hobbes’s state of nature, where the wholesale pursuit of self-interest leads everyone to a condition in which he or she is worse off. In such a scenario, argues Hobbes, it is in every man’s interest to endeavour peace with his neighbour, and therefore:

---

101 Darwall, p. 2 (original emphasis).
102 Darwall, p. 2 (original emphasis).
“…be willing, when others are so too, as farre-forth, as for Peace, and defence of himself he shall think it necessary, to lay down this right to all things; and be contented with so much liberty against other men, as he would allow other men against himself.”

In this way, Hobbes argues that self-interested rationality should lead individuals to create bargains in which “each accepts certain constraints on his freedom of action so that all avoid the costs of the natural condition of war.” I would, for example, have good reason to agree not to kill my neighbour, on the understanding that he will agree not to kill me. Likewise, I might make similar agreements to respect his demarcated lot of land and private possessions on the understanding that the same respect is shown towards my own property.

This is how rational beings can avoid the state of nature—and, incidentally, how the individuals in the prisoner’s dilemma can avoid a long period of incarceration: cooperation. Individuals will cooperate when they “forego the pursuit of their own independent interests” and follow rules which, when followed collectively, promote “everyone’s interests better than would have been done by everyone pursuing her own interests independently.”

Contractarians then use Hobbes’ observation as a springboard for a full ethical theory:

Like Thomas Hobbes, the contractarians begin by assuming that the parties to the social contract are rational agents seeking to advance their interests; whereas Hobbes argued that the foundation of political authority is in the social contract, however, the contractarians propose to derive the

---

104 Hobbes, Part I, Chapter 14, p. 190.
107 Darwall, p. 3.
108 Darwall, p. 3.
fundamental terms of social life—that is, *morality*—from the contract. Hence the title of what is probably the most influential work in this category, David Gauthier’s *Morals by Agreement*.

Contractarianism, then, starts with Hobbes’ assumption that “the parties to the social contract are rational agents seeking to advance their interests,”¹¹⁰ then combines this with the further claim that “a rational assessment of the best strategy for *attaining* the maximization of their self-interest will lead them to act morally (where the moral norms are determined by the maximization of joint interest).”¹¹¹ So, while contractarianism may begin from the observation that rational agents are *motivated* by the advancement of their own interests, it does not see these individuals as simply promoting those interests. Instead, individuals cooperate—where cooperation promotes everyone’s advantage, but does so “by requiring individuals to *restrict* the promotion of their interests.”¹¹²

**Distinguishing Contractarianism from Contractualism**

The next step is to apply this contractarian theory to the specific problem of punishment. Before doing this, however, it may be helpful to clear up a potential confusion in the terminology I am using. So far, I have used the term ‘contractarianism’ to refer to the specific theory proposed by Thomas Hobbes and later developed by David Gauthier. Within the wider social contract

¹¹⁰ Dagger, p. 344. While contractarian agents are here—and in most texts—referred to as maximising their own “interests,” it is perhaps more accurate to talk of them being focussed on their own *goals*—the objects of which, on occasion, may very well be the interests of others. Consider, for example, the doting father who wants nothing more than to see his daughter achieve her dream of becoming a philosophy professor, or the dedicated paramedic who wants only to save as many lives as she can.

¹¹¹ Ann Cudd, “Contractarianism,” in *The Stanford Encyclopedia of Philosophy* <http://plato.stanford.edu/entries/contractarianism/> (emphasis added). This claim refers specifically to the community-wide cooperation that occurs within a society. It does not necessarily suggest that a more isolated case of maximised-joint-interest (such as the two prisoners keeping silent) will necessarily count as moral behaviour.

¹¹² Darwall, p. 4.
tradition, however, there is a competing alternative—namely, the view argued for by Jean-Jacques Rousseau, John Rawls, and Timothy Scanlon, and commonly known as contractualism.

Both views hold that individuals in a condition of nature will choose to come together to form a state, but differ in how they argue for this conclusion. Like contractarianism, contractualism also understands principles of right conduct as the object of a rational agreement.\textsuperscript{113} However, while contractarianism sees moral principles as arising from some specific system of rationally self-interested bargaining,\textsuperscript{114} contractualism instead sees the relevant agreement as arising from moral ideals of respect,\textsuperscript{115} reasonable reciprocity, and fairness between moral equals.\textsuperscript{116} Contractualism begins with the reasonable individual who is already engaged in social cooperation, and who wants to find and live according to fair or just principles in reciprocity with others.\textsuperscript{117} According to this view, morality consists in “what would result if we were to make binding agreements from a point of view that respects our equal moral importance as rational autonomous agents.”\textsuperscript{118} This can be contrasted with contractarianism, where the parties’ equality “is merely \textit{de facto} and their choice of principles rationally self-interested.”\textsuperscript{119}

The distinction between contractarianism and contractualism might therefore be summarised as follows: While contractarianism claims that I seek to maximise my own interests in a bargain with others, contractualism instead claims that I seek to pursue my interests in a way that I can justify to others who are also pursuing their own interests.\textsuperscript{120}

\textsuperscript{113} Darwall, p. 4.  
\textsuperscript{114} This bargaining will more often than not be hypothetical—not actual. More on this in Chapter 4.  
\textsuperscript{115} Darwall, p. 4.  
\textsuperscript{116} Darwall, p. 1 (original emphasis).  
\textsuperscript{117} Dagger (2011), p. 345 (emphasis added).  
\textsuperscript{119} Darwall, p. 1 (original emphasis).  
\textsuperscript{120} Ashford and Mulgan.
To confuse matters, some authors use the term ‘contractualism’ in a much broader sense to refer to all theories within the social contract tradition—including the narrow version of contractualism just defined, and the contractarian approach outlined in the previous section.\textsuperscript{121} To worsen matters, other authors use the term ‘contractarianism’ in this same broad way. Claire Finkelstein, for example, sketches the history of social contract theory by claiming that: \textsuperscript{122}

There is a robust contractarian tradition that emerged in seventeenth century political philosophy, first with the writings of Thomas Hobbes, later in the Enlightenment version of this same tradition in the writings of Locke and Rousseau, and finally in a Kantian version of the tradition, as developed by John Rawls.

In this way, Finkelstein uses the term ‘contractarianism’ to refer to the wider tradition of social contract theory that encompasses both of the contractarian and contractualist theories discussed above. In order to distinguish between these two narrower approaches, Finkelstein goes on to employ the terms ‘normative contractarianism’ and ‘rational contractarianism’ in lieu of ‘contractualism’ and ‘contractarianism’ respectively.\textsuperscript{123}

To be clear, then, when I use the term ‘contractarianism’, I refer specifically to the Hobbesian theory outlined earlier in this chapter—not the wider social contract tradition that also encompasses what I have here identified as ‘contractualism’.\textsuperscript{124}

\textsuperscript{121} Ashford and Mulgan.
\textsuperscript{122} Finkelstein (2011), p. 322.
\textsuperscript{123} Finkelstein (2011).
\textsuperscript{124} Hon-Lam Li has recently defended a contractualist theory of punishment (see Hon-Lam Li, “Contractualism and Punishment,” Criminal Justice Ethics, Vol. 34, No. 2 (2015), pp. 177-209), though it differs from my own contractarian theory of punishment in exactly the conceptual ways just described.
A Contractarian Theory of Punishment

The remainder of this thesis is dedicated to arguing that contractarianism is well-suited to the task of providing a moral justification for punishment. The fundamental claim underpinning my argument will be that agreement carries some kind of special normative force. It is worth noting, however, that “there is nothing altogether new in the claim that criminals somehow have authorized or consented to their punishment.” Indeed, as Richard Dagger notes:

The social contract theorists of the seventeenth and eighteenth centuries advanced such a claim, each in his own way, with Jean Jacques Rousseau stating the point most vividly: “it is in order not to be the victim of a murderer that a person consents to dies if he becomes one.” Rousseau’s reasoning, however, strikes many readers as counterintuitive. To say that a murderer deserves to die for his or her crime is controversial in itself; to say that the murderer has consented to be executed seems not only to defy experience but to stretch the concept further than it can reasonably go. That, presumably, is why the vigorous debates over the justification of criminal punishment in the twentieth century had so little to say about the putative connection between the offenders’ consent and their punishment.

However, while consent has often featured in theories of punishment, Claire Finkelstein notes that there has been an absence of a specifically contractarian alternative to consequentialist and retributivist theories of punishment—an absence that is, according to her, “especially

---

125 Dagger, p. 341.
126 Dagger, p. 341.
surprising in view of the breadth and depth of the contractarian school of thought in political theory.”128

On occasion, we have seen arguments that seem to suggest something like a contractarian justification for punishment—but these remain largely undeveloped. As an example, Finkelstein recounts the discussion that Socrates has with his student, Crito, regarding the possibility of Socrates escaping from prison and avoiding his own execution. He refuses to take such an opportunity as he sees himself as bound to submit to his sentence, despite believing that it is a gross miscarriage of justice. Socrates’s argument for this position is—as Finkelstein summarises it—that he had “entered into an agreement with the State to abide by its laws, in exchange for which he has enjoyed all the benefits of Athenian citizenship.”129 By giving this agreement, argues Socrates, he is now bound to subject himself to the punishment that has been handed down to him.

Even Finkelstein’s own contractarian argument for punishment only goes so far as attempting to justify punishment as a general institution. I intend to take a far more detailed approach, arguing (via contractarianism) for the moral permissibility of specific instances of punishment, and—from this—for the moral permissibility of more general society-spanning punishment practices.

My argument begins from what I take to be a fairly uncontroversial claim—that is: that for every person \( P \), there are certain things that are valuable for \( P \). My theory is largely indifferent to the precise source of this value. What is valuable for an individual may be objective (that is, valuable independent of her attitudes towards those things) or subjective (that is, valuable precisely because of that individual’s judgements, desires, or what brings her

---

pleasure). If successful, my theory is equally capable of operating under either of these conceptions of value. For the remainder of this thesis, however, I will assume a subjectivist approach to value. My reason for doing this is that—for reasons that will become apparent in Chapter 6—a contractarian theory of punishment built on a subjectivist conception of value is more problematic than one built on objectivism. Thus, if I can demonstrate the workability of the former, I can also assume the workability of the latter.

Provided that we accept the above claim, it seems that we must also accept the corollary claim that if there are certain things that are valuable for each individual, then there are certain kinds of behaviour which each individual has reason to disvalue. Suppose we find that ‘continued survival’ is valuable for a person. If this is the case, then that person has good reason to disvalue behaviours which put her life at significant risk. Likewise, if it turns out that the unmolested possession of private property is valuable for a person, then that person will have reason to disvalue behaviour which might result in the theft or damage of what she owns.

If an individual has reason to disvalue certain kinds of behaviour, then it seems that she will also have good reason to deter those very same behaviours. This is where the role of punishment comes into play. Punishment can be seen as a system through which individuals are able to ‘purchase’ deterrence so as to prevent others from engaging in certain kinds of behaviour. So, if a particular punishment practice brings with it a “moderately strong deterrent efficacy” against behaviours that we have reason to disvalue, then a rational contractor will, according to Finkelstein, “agree to live in a regime that furnishes this level of deterrence… and hence would prefer it to one in which deterrence is absent.”

This reasoning can be summarised as follows:

---

1. For every person $P$, there are certain things that are valuable for $P$.

2. If there are certain things that are valuable for $P$, then $P$ has good reason to disvalue certain kinds of behaviour.

3. If $P$ has good reason to disvalue certain kinds of behaviour, then $P$ has good reason to deter these same behaviours.

4. If $P$ has good reason to deter certain kinds of behaviours, then $P$ has some reason to agree to the implementation of punishment practices that effectively achieve this deterrence.

Therefore,

5. $P$ has some reason to agree to the implementation of some kinds of punishment practices.

Of course, there is an implied premise in this argument: namely, the empirical claim that punishment is capable of creating a deterrent effect. One way to object to the contractarian theory outlined above would be to deny that such a relationship holds—that is, to deny that punishment ever deters. It is important to note, however, that the contractarian theory of punishment requires only the most minimal claim regarding deterrence in order to get off the ground. It does not require that all forms of punishment are capable of deterrence, nor that any of them are particularly good forms of deterrence. All that is required is that there are some conceivable punishments that are capable of some deterrence. So long as this is the case, the contractarian theory of punishment will make it out of the starting gate.

What appears above, then, is the basic structure of the contractarian theory of punishment. So long as there are things which are valuable for an individual (be the source of
this value objective or subjective) the individual will have reason to agree to certain kinds of
punishment practices. This agreement—once given—is what grounds the moral justification of
punishing that individual.

Finkelstein notes that it is typically easier to justify the enforcement of voluntary
arrangements over involuntary ones, and that if this is the case “then it is clear that a theory of
punishment that convincingly predicates a consensual foundation for the institution should depict
the institution as easier to justify than other types of theories.”¹³² A contractarian theory of
punishment does exactly this, uniquely combining “the social aim of deterrence with an
individualised approach to the justification for imposing punishment on a particular agent, thus
providing the criminal with an argument for his own punishment that he can accept.”¹³³

But can the moral permissibility of a punishment practice be derived merely from the
consent of the individuals to whom it applies? Finkelstein thinks not, claiming that “consent
considered by itself does not have such normative force, as is reflected in the fact that the
criminal law rejects consent as a defence to most crimes.”¹³⁴ On first glance it may seem that this
claim is obviously false. It certainly appears as though consent can absolve an offender of
wrongdoing in a number of scenarios. The presence of consent can, for example, turn theft into a
gift, battery and assault into competitive sport (boxing), and murder into active voluntary
euthanasia.¹³⁵

Perhaps consent is all that is needed. If it is, then the evidential burden for the
contractarian will be minimal. In order to demonstrate the moral permissibility of a harmful
action, they will only need to point to the prior agreement of the victim. Nevertheless, I share

¹³⁵ My thanks go to Professors Baruch Brody and George Sher for providing me with these examples.
Finkelstein’s scepticism that consent is enough—and a deeper examination of the examples given in the previous paragraph fuels this doubt. The importance of consent in the case of theft is not as simple as suggested above. In several cases, the English courts have held that theft may occur notwithstanding the consent of the owner.\textsuperscript{136} The circumstances in which consent can be a defence to assault and battery are also severely limited. Outside of certain minor exceptions, consent can never operate as a defence to the deliberate infliction of bodily harm.\textsuperscript{137} Participants in a sports game are treated as consenting to the possible bodily harm that is an essential element of their sport,\textsuperscript{138} but this can never amount to actual or grievous bodily harm\textsuperscript{139} (here, boxing is an anomalous exception for a variety of historical reasons). Further, while the moral permissibility of active voluntary euthanasia remains an open question, in many jurisdictions—such as New Zealand—it remains the case that:\textsuperscript{140}

\begin{quote}
[n]o one has a right to consent to the infliction of death upon himself or herself; and, if any person is killed, the fact that he or she gave any such consent shall not affect the criminal responsibility of any person who is a party to the killing.
\end{quote}

Indeed, even in jurisdictions which have legalised euthanasia, mere consent is insufficient. Further requirements—such as the patient suffering from a terminal disease—are almost always in place.\textsuperscript{141} This provides further evidence that consent taken alone provides insufficient normative force to turn an illegal / morally impermissible action (murder) into a legal / morally permissible one (active voluntary euthanasia).

\textsuperscript{136} See \textit{Lawrence v MPC} [1972] AC 626 HL and \textit{DPP v Gomez} [1993] AC 442 HL.
\textsuperscript{137} See \textit{R v Brown} [1993] UKHL 19.
\textsuperscript{138} See \textit{R v Johnson} (1986) CA.
\textsuperscript{139} See \textit{R v Cato} [1976] 1 WLR 110 CA.
\textsuperscript{140} The Crimes Act 1961 (New Zealand), s. 63.
\textsuperscript{141} See, for example, the conditions contained within California’s recent ‘end of life’ bill <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201520160SB128>.
Thus, while some minor exceptions occur, it remains largely true that—per Finkelstein—agreement alone lacks the normative force to make harmful actions morally permissible. It is for this reason that Finkelstein suggests we look to the Socratic notion that “it is not consent alone that justifies punishment, but consent premised on the benefit the citizen who consents to abide by the State’s dictates takes himself to be receiving in the bargain.”142 Put simply then, the moral permissibility of a particular punishment practice will depend on both an individual’s agreement to the implementation of that punishment practice, and the individual having good reason to agree to that practice—say, because she will derive some kind of benefit from it.

This is not merely an ad hoc move—we have a principled reason for turning to benefit as an additional criterion for the test. Specifically, objections to consent as a justifier focus on cases of consensual harm that still seem to be bad for the one who is harmed. By including a benefit requirement, we answer such objections in the very same currency, saying something like “while this particular experience might be harmful for the punishee, she is in fact benefitted on the whole.” It is unclear how burdensome this criterion should be, however—specifically on whether there must there be a guarantee of benefit (i.e., an assurance that the individual will actually benefit from the implementation of the punishment practice) or merely an expected benefit (i.e., the individual stands a better chance of benefitting in a world with the punishment practice than without it). I will discuss both alternatives in Chapter 3, before settling on the latter as the most appropriate candidate. Whichever standard we adopt, it must be sufficient to give the individual good reason to become a party to the agreement. For this reason, we will temporarily strip back this criterion to the bare claim that $P$ must have “good reason to agree.”

---

To reiterate, then, our contractarian theory of punishment *in utero* holds that it will be morally permissible to punish an individual $P$ under a particular punishment practice $X$ so long as, at some prior time, $P$ simultaneously:

(i) Had good reason to agree to $X$; and

(ii) Gave some kind of agreement to $X$.

Of course, it still remains to be seen what form the agreement that underpins the contractarian theory *actually takes*. I will ultimately argue that (i) and (ii) are linked. Specifically, I will contend that the latter is derived from the former. I diverge from Finkelstein’s approach in that I do not believe that we can look to any specific historical agreement in order to discover a moral justification for particular punishment practices. Instead—as I will argue in Chapter 4—we must look to some kind of hypothetical agreement. I will argue that an individual can be taken to hypothetically agree to a particular punishment practice where it can be shown that he or she can expect to benefit from that practice, and thus would have good reason to agree to its implementation.
Chapter 3: How We Have Good Reason to Agree to Punishment

In the previous chapter, I proposed that it will be morally permissible to punish an individual $P$ under a particular punishment practice $X$ so long as, at some prior time, $P$ simultaneously:

(i) Had good reason to agree to $X$; and

(ii) Gave some kind of agreement to $X$.

I will turn to consider exactly how an individual agrees to a punishment practice in the next chapter. For now, however, I intend to focus on the first of these two criteria. Specifically, I will attempt to answer the following question: In what circumstances will an individual have good reason to agree to a particular punishment practice?

We began our sketch of the contractarian theory by considering what incentive an individual might ever have to agree to the imposition of a punishment practice. The answer, we decided, was the opportunity to purchase deterrence. The fact that a punishment practice is capable of deterring behaviour that an individual disvalues will provide prima facie evidence for her having some reason to agree to that practice.

Imagine, for example, that $P$ lives in a society in which there is no punishment for theft. $P$ calculates that she suffers an average loss of $1000$ a year from having her property stolen. Further, $P$ strongly disvalues this loss. Suppose, then, that a particular punishment practice for theft is proposed—and that evidence is provided that the implementation of this practice will strongly deter (and therefore significantly reduce) instances of theft. In fact, it quickly becomes clear that with the implementation of this practice, $P$ will only stand to suffer an average loss of
$100 a year from having her property stolen. Such an observation will—taken alone—provide $P$ with good reason to agree to the implementation of the punishment practice in question.

But deterrence does not come free. Agreeing to the implementation of a punishment practice may create utility in the form of deterrence, but it also brings with it the possibility of disutility. In the following section I will identify three potential sources of disutility within a punishment practice: what I shall refer to as Sanction, Restraint, and Collateral.

**The Disutilities of Sanction, Restraint, and Collateral**

Most of us will manage to avoid many punishments. In fact, many of us will live our lives avoiding punishment altogether. Nevertheless, some individuals will suffer punishment—either because they are responsible for committing an offence, or because they are mistakenly identified as an offender. In most cases, this punishment will also be strongly disvalued by the individual who is punished (hereafter referred to as the ‘punishee’). The reason for this is that punishment is—by the very definition given in Chapter 1—necessarily harmful.

All this is to say that there is, for each of us, some non-zero probability that we will at some point find ourselves punished under a particular punishment practice. When this occurs, we will presumably suffer some detriment as a result of the harm that is inflicted upon us. This is the disutility of Sanction—hereafter referred to as $U_{\text{Sanction}}$. The actual value of this disutility will depend on several variables. Two obvious factors are the kind and severity of the punishment in question. But depending on the theory of value we adopt, other more subjective elements may come into play. Certain punishments may hold varying disutilities for different individuals. The disutility of a fine may, for example, differ between someone who is destitute and someone who
is well-off. The disutility of a prison sentence may also differ depending on how well-adjusted an inmate is. Even the disutility of more extreme punishments—like the death sentence—may vary depending on the amount of value that the punishee attaches to his or her continued existence.

Even if an individual is not punished, she may suffer disutility in another form. It is true that by agreeing to the implementation of a punishment practice, an individual creates a way of deterring others from committing a particular crime. But that deterrence also extends to the very individual who gave her agreement. Put simply, the implementation of the punishment practice may require the individual to now restrain herself from certain beneficial behaviour that she might otherwise have engaged in. There are many things that we might disvalue others doing, but which are in fact very valuable for us. We might, for example, disvalue other drivers speeding because of the unnecessary risk it poses to our own lives. All the same, we might strongly value our ability to speed in order to save time and avoid being late to meetings. Likewise, while we might disvalue theft when perpetrated by others, we stand to gain a lot if we are personally able to engage in stealing the property of others (provided, of course, that we can do so undetected). Nevertheless, the implementation of a punishment practice may cause us to forego such behaviours—to obey laws related to speeding and theft for fear of being caught and punished. This is the disutility of Restraint, hereafter referred to as $U_{\text{Restraint}}$.

Like $U_{\text{Sanction}}$, the value of $U_{\text{Restraint}}$ will vary on a case-by-case basis. For many, the inconvenience of refraining from certain offences will be minor, while for others this burden will be particularly onerous. Ultimately, the disutility of restraint will depend on two central factors: (1) the value for the individual of committing the offence, and (2) the inclination of the individual to commit the offence in the first place. Suppose you discovered that you were the sole heir to your uncle’s enormous fortune. Clearly, you have much to gain by murdering him.
But for many of us, carrying out the murder is something we will have no inclination to do in the first place. For those of us who feel this way, restraining ourselves from murder will carry minimal disutility. If, on the other hand, we would be inclined to carry out the murder, then the existence of a punishment practice that sufficiently deters us from carrying out the killing will cause us greater disutility in the form of restraint.

In addition to $U_{Sanction}$ and $U_{Restraint}$, there are a number of other **Collateral** disutilities associated with the implementation of a punishment practice—hereafter referred to collectively as $U_{Collateral}$. One example of such a disutility is the practical cost of implementing a punishment practice. Effective punishment requires detection, investigation, adjudication, and administration—none of which comes free. Consider again the example from above where the implementation of a particular punishment practice managed to reduce $P$’s average yearly loss resulting from theft from $1000$ to $100$. Suppose, however, that the overwhelming expense of this practice required every citizen—including $P$—to pay an additional $2000$ in annual taxation. In such a case, the cost of the punishment practice would grossly outweigh the benefit gained from deterrence.

A further case of collateral disutility is the potential for a punishment practice to incentivise the commission of other crimes. Justice Kennedy, for example, has argued that increasing the penalty for rape to death would decrease the disincentive for offenders to murder victims of rape.\footnote{See *Kennedy v Louisiana*, 554 U.S. 407, 445 (2008).} Finkelstein provides another example, noting that if the penalty for bicycle theft is increased too far, this may encourage potential offenders to instead steal automobiles.\footnote{Finkelstein (2011), p. 333.} If it can be shown that a particular punishment agreement may incentivise the commission of other (especially more serious) crimes, then this will be a collateral disutility that must be
weighed against the utility of deterrence when considering whether or not the punishment practice should be implemented.

Having discussed these three potential sources of disutility, we can now return to the question with which we began this chapter: In what circumstances will an individual have good reason to agree to a particular punishment practice?

To Agree or Not to Agree

When faced with a proposed punishment practice, an individual will have one of two options: to agree to the practice, or to not agree to the practice. Depending on which way an individual decides, there are three potential future outcomes for the individual in question.

First, there is the scenario in which the individual refuses to agree to the punishment practice, and the practice is—as a result—not implemented. In this scenario, nothing will change, and the total utility value gained by the individual will remain the same as that prior to considering the implementation of the practice. I will call this outcome Status Quo.

If, on the other hand, the individual agrees to the punishment practice in question, there are two possible scenarios that may come to pass. Upon the individual’s agreement, the punishment practice will be implemented—conferring on her all of the benefits of deterrence that come with that practice. The individual may then go an entire lifetime without being punished under the practice. I will call this outcome Beneficial Punishment. On the other hand, there is—as noted earlier—some non-zero probability that the individual will, in fact, find herself punished under the very practice she has agreed to. I will call this outcome Detrimental Punishment. These three outcomes can be presented visually as follows:
In order to establish what \( P \) has \textit{good reason to do}, we need to attach utility values to each of these three potential outcomes. The utility of Status Quo is the utility value that will be rendered for the individual if no punishment is implemented for the offence in question. I will refer to this utility value as \( U_{\text{Status Quo}} \). Analysing the expected utilities of both Beneficial Punishment and Detrimental Punishment is more complex. In order to do this we must return to the collection of utilities and disutilities discussed earlier.

As noted in Chapter 2, the implementation of an effective punishment practice will result in a quantifiable increase in utility for an individual—namely, the \textit{Deterrence} of certain kinds of behaviour that she disvalues. I will refer to this gained utility as \( U_{\text{Deterrence}} \). The actual value of \( U_{\text{Deterrence}} \) depends on two key factors. The first—and most obvious—of these will be the harmfulness of the deterred offence. An individual has much more to gain from deterring an offence that causes a great deal of harm (like murder) than one which causes only a minor amount of harm (like petty theft). The second relevant factor is the \textit{amount} of deterrence purchased by the implementation of the punishment practice. It is not enough to simply say that a particular punishment practice \( X \) deters the commission a particular crime \( C \). The \textit{extent} to which
C is deterred by $X$ must also be considered. These two factors—the harmfulness of the crime, and the extent to which the crime is deterred—will then need to be considered in tandem. A modest increase in the deterrence of an incredibly harmful offence may, for example, render a higher $U_{\text{Deterrence}}$ than a massive increase in the deterrence of a less harmful offence. The value of $U_{\text{Deterrence}}$ will be the same for both Beneficial and Detrimental Punishment. That is, the utility gained in the form of deterrence will be the same regardless of whether or not that individual goes on to be punished under that particular practice. Where these two outcomes differ is in the other disutilities that each attracts.

Consider, first, Beneficial Punishment—that is, the outcome in which the punishment practice is implemented, and the individual never suffers the punishment in question. An individual who experiences this outcome will receive the utility of $U_{\text{Deterrence}}$, but may also be subject to the disutilities of $U_{\text{Restraint}}$ (restraining herself from certain behaviours that may now—thanks to the implementation of the punishment practice—put her at risk of being punished) along with the disutilities of $U_{\text{Collateral}}$ (the cost of the practice etc.). The total utility of Beneficial Punishment can therefore be given by the following sum:

$$U_{\text{Beneficial Punishment}} = U_{\text{Deterrence}} + U_{\text{Restraint}} + U_{\text{Collateral}}$$

That is, the value of Beneficial Punishment for an individual will be equal to the total utility she gains from the deterrent effect of the punishment practice combined with the total disutility she suffers in restraining herself from engaging in the prohibited behaviour and any other collateral disutilities.
Consider, next, the utility value of Detrimental Punishment—that is, the outcome in which the punishment practice is implemented, and the individual does suffer the punishment in question. Once again, this individual will receive the utility of $U_{\text{Deterrence}}$. This time, however, she will also receive the disutility of $U_{\text{Sanction}}$ (some particular disutility as a result of being punished under the punishment practice) along with the associated disutility of $U_{\text{Collateral}}$. Somewhat counter-intuitively, she may also receive some of the disutility of $U_{\text{Restraint}}$. There are two reasons for this. First, while an offender may fail to restrain herself in one instance, she may have shown restraint—and therefore suffered a disutility—in other circumstances. Second, some individuals may be punished incorrectly, and thus have never committed a crime (and therefore always shown restraint. The total utility of Detrimental Punishment can therefore be given by the following sum:

$$U_{\text{DetrimentalPunishment}} = U_{\text{Deterrence}} + U_{\text{Sanction}} + U_{\text{Restraint}} + U_{\text{Collateral}}$$

That is, the value of Detrimental Punishment will be equal to the total utility she gains from the deterrence effect of the punishment practice combined with the total disutility she suffers in being subjected to the punishment in question and any other collateral disutilities.

How, then, can we use the utility values of these three potential outcomes (Status Quo, Beneficial Punishment, and Detrimental Punishment) to establish in what circumstances an individual will have good reason to agree to a particular punishment practice? A difficulty lies in the fact that both Beneficial Punishment and Detrimental Punishment are outcomes of the same decision—that is, the decision to agree to the implementation of the punishment practice. Beneficial Punishment should always render a utility higher than the Status Quo (for, if it didn’t,
there would be no reason to even entertain the implementation of the punishment practice in the first place). Detrimental Punishment, on the other hand, may render a lower utility value than the Status Quo—particularly for those with a predilection for offending. For these individuals, the implementation of a punishment practice will therefore be a gamble. Whether or not the practice leaves them in a better or worse position than the status quo will depend on a number of future factors. Given this, how are we to accurately weigh the benefits of implementation versus non-implementation?

**Guaranteed Benefit**

One way would be to assign a probability to each of Beneficial Punishment and Detrimental Punishment—that is, to estimate the likelihood that an individual will, at any time in the future, be subjected to the punishment under consideration. With this probability in hand, we could then calculate the average utility value of implementation and weigh this against the utility value of remaining with the status quo. Put another way, the utility of agreeing to the punishment practice would be given by (where \( p \) is the probability of being punished for the offence in question):

\[
(U_{\text{Detrimental Punishment}} \times p) + (U_{\text{Beneficial Punishment}} \times [1-p])
\]

An alternative—and much simpler—approach would be to instead focus solely on the worst of the two outcomes: that is, Detrimental Punishment. As Finkelstein notes, individuals tend to “eschew gambles where the basic elements of their well-being are concerned”\(^{145}\) and instead

“seek assurances that they will benefit under any future life circumstance or choice they might make.”¹⁴⁶ In this way.¹⁴⁷

Rational agents entering into agreements for basic social institutions would require… that institutions to which they give their assent would actually improve their conditions, as compared with the lives they would lead in their absence.

Put another way: when it comes to agreements pertaining to the basic structure of society, individuals remain agnostic about their future choices and seek a promise of benefit for the worst case scenario.¹⁴⁸ Finkelstein calls this the ‘benefit principle’ and notes a similar concept operating in the works of both John Locke¹⁴⁹ and John Rawls.¹⁵⁰ As Finkelstein notes, the benefit principle would be far too strong as a general condition of rationality, ruling out many cases of insurance, gambling, and stock market investment.¹⁵¹ The claim we are making here, however, is that—per Locke, Rawls, and Finkelstein—such a strong condition is not irrational when it comes to decisions regarding the basic structure of society. The reason for this is simple. Individuals are assumed to be agreeing to these societal structures ex ante—that is, prior to having full knowledge of how their lives will turn out. Given this uncertainty, individuals will be highly motivated to only agree to societal structures that will benefit them regardless of what future decisions they make, or what kinds of circumstances befall them.

Consider, then, how this benefit principle can be applied to the specific problem of punishment. As Finkelstein notes, it seems that the threat of punishment is “quite essential to a

¹⁴⁷ Finkelstein (2005), p. 215 (original emphasis).
social order predicated on voluntary agreement.”\textsuperscript{152} This is because a contract will only be effective where there is \textit{compliance}. The parties to a social contract must therefore have “some way of ensuring continued compliance with the terms of the agreement, given the temptation members will have to offer their initial consent and then free-ride on the compliance of others while silently defecting.”\textsuperscript{153} The social contract must set out consequences for violators and “a plausible enforcement mechanism for detecting violations and imposing the announced penalties.”\textsuperscript{154}

In this way, a system of punishment will be a necessary part of any agreement that sets out substantive rules of compliance.\textsuperscript{155} Thus, on the benefit principle, we will have to ask whether each member of society “will regard himself as faring better, under an institution that mandates punishment, than he would fare in the absence of such an institution.”\textsuperscript{156} In order to establish this:

Each member of society must project himself into the position of someone who has violated the conditions of the more basic, substantive social contract, and ask himself whether, if he were to be punished for such violations, \textit{he would still fare better than he would had he never agreed to live under threat of punishment in the first place}.\textsuperscript{157}

An individual will therefore have good reason to agree to the implementation of a particular punishment practice (thus making it morally permissible to punish her under that practice) so long as she still benefits even in the worst case scenario—that is, the scenario where she herself

\textsuperscript{152} Finkelstein (2005), p. 215.
\textsuperscript{153} Finkelstein (2005), p. 215.
\textsuperscript{154} Finkelstein (2005), p. 215.
\textsuperscript{155} Finkelstein (2005), p. 215.
\textsuperscript{156} Finkelstein (2005), p. 215.
\textsuperscript{157} Finkelstein (2005), p. 215-216 (emphasis added).
receives the punishment in question. According to the Benefit Principle, then, an individual will have reason to agree to a punishment practice where the utility values of Beneficial Punishment and Detrimental Punishment\(^{158}\) are both greater than the utility value of the Status Quo. That is, where:

\[
U_{\text{Detrimental Punishment}} > U_{\text{Status Quo}} < U_{\text{Beneficial Punishment}}
\]

Or, substituting in the utility values from earlier, where:

\[
[U_{\text{Deterrence}} + U_{\text{Sanction}} + U_{\text{Restraint}} + U_{\text{Collateral}}] > U_{\text{Status Quo}} < [U_{\text{Deterrence}} + U_{\text{Restraint}} + U_{\text{Collateral}}]
\]

This formula provides us with a precise test for when an individual will (1) be guaranteed to benefit from a particular punishment practice, and therefore (2) have good reason to agree to the implementation of that practice. By following this approach, the contractarian theory is able to generate normative constraints on punishment\(^{159}\) and “dictate specific parameters for the punishment of each separate crime.”\(^{160}\)

Finkelstein is optimistic about the possibility of the benefit principle providing moral justification for a wide number of useful punishment practices. She argues that: \(^{161}\)

\(^{158}\) In this way, ‘Detrimental Punishment’ may be something of a misnomer, as this outcome will—in some circumstances—be better for the individual than the status quo.


\(^{161}\) Finkelstein (2005), p. 216.
…for many sanctions the benefit test will be satisfied [because] a complete absence of any form of punishment for violations of the social contract would eliminate the possibility of social cooperation entirely, since an agreement would unravel without threat of enforcement.

She goes on to claim that, because of this: 162

…for most penalties, and most societies, even an offender who must suffer punitive sanctions will fare better under a punishment agreement than he would in the absence of all social enforcement.

But it is important to remind ourselves precisely what the baseline for comparison is when considering whether an individual is guaranteed to benefit under a punishment practice. My intention is not—contra Finkelstein—to compare a world in which the practice is implemented with a world in which there is an absence of all social enforcement. This standard seems too easy to meet. Instead, each individual is merely expected to compare the world in which the punishment practice in question is implemented with the world in which that specific practice is not (the Status Quo). If the individual is guaranteed to benefit by choosing the former, then it will be morally permissible to implement that punishment practice over that individual.

None of this is to say that the benefit principle will uniquely identify the best punishment practice available. Rather, the claim is simply that rational contracting agents would reject any punishment practice that fails to meet this standard. 163 It will almost certainly be the case that, for each individual, there will be a range of punishment for a specific crime that will benefit her more than the status quo. In truth, she will have reason to agree to any of these, and will most likely prefer the punishment practice that has the best worst-case scenario (that is, the highest

value for $U_{\text{DetrimentalPunishment}}$). For now, however, I will leave aside the question of how a specific punishment practice should be chosen from a range of morally permissible alternatives. This is because it will later become apparent that such a decision must be made in tandem with the punishment practice preferences of other individuals within a society. I will turn to discuss this properly in Chapter 5.

**A Lower Standard than Guaranteed Benefit?**

While I have established guaranteed benefit as being sufficient for an individual to have good reason to agree to a punishment practice, this does not necessarily mean that it is necessary. My claim so far has merely been that if anything will provide an individual with good reason to agree to a punishment practice, guaranteed benefit certainly will.

It is important, then, to consider whether a lower standard than guaranteed benefit may still provide a sufficient foundation for the contractarian theory of punishment. We might, for example, consider whether mere expected benefit will suffice. Put another way, we can explore the claim that an individual will still have good reason to agree to a punishment practice where she only expects to benefit from that practice, not where such benefit is automatically guaranteed.

The principle concern with adopting this lower threshold might be that we move from a benefit of which we are certain, to a benefit which is—to some extent—a gamble. While the former approach assures us that an individual will benefit from the implementation of the practice, the latter approach can provide no such certainty. The individual may very well turn out to be worse off under the practice, even if she expected to fare better. While this concern does
have some merit, it is important to clarify the exact loss of certitude that might occur by reducing our standard from guaranteed benefit to expected benefit. For one, the ‘certainty’ contained within guaranteed benefit is not as certain as it might first appear. The principle source of benefit—the very thing which has the potential to make Detrimental Punishment preferable to the Status Quo in the first place—is the increase in deterrence created by the punishment practice. But it will be rare that a punishment practice is ever able to fully eradicate all instances of a particular crime. Instead, the benefit gained from deterrence comes in the form of a reduction in the rate of offending—and therefore a reduction in the likelihood that an individual will be a victim of the crime in question. The practice may bring about substantial deterrence—and therefore a substantial reduction in the rate of offending—but if the chances of becoming the victim of a crime are probabilistic, then there will still be some possibility of an individual being victimised even after agreeing to a punishment practice for that crime.

Thus, when referring to a ‘guaranteed’ benefit, we are not suggesting that an individual is guaranteed to avoid being the victim of a particular crime. There is still some possibility that such victimisation may occur. But an individual does not have to be guaranteed immunity against all future offending in order to be guaranteed a benefit. Instead, the sense of benefit we are using here suggests that an individual fares better whenever the probability of victimisation is reduced—even if the individual goes on to nevertheless be a victim of offending. This benefit will come in two primary forms.

First, there is the reduction in fear and anxiety that might accompany a sufficient reduction in offending.\(^{164}\) The more likely we are to be victimised (and the more aware we are of this likelihood) the more our cognitive faculties will be taken up by negative attitudes and dispositions towards this eventuality. When the likelihood of victimisation is reduced, however,

---

\(^{164}\) Thanks to Professor Donald Morrison for suggesting this.
the need for these negative mental states is reduced. This is a benefit which is guaranteed regardless of whether or not an individual actually goes on to be victimised.

Second, an individual will fare better merely by the fact that she has better odds of avoiding victimisation. This claim, I contend, is an intuitive one. Suppose that I play a card-based game of chance in which my odds of winning are one-in-ten, or 10%. Suppose, then, that a friend teaches me a clever trick which improves my odds of winning to one-in-four, or 25%. Does my friend’s advice benefit me? Clearly, the answer to this question is ‘yes’. My friend’s advice greatly increases by chance of winning, and thus clearly benefits me when playing the game. Most importantly, this benefit does not necessarily depend on my actually winning the next game. I may very well lose. But if I do, the benefit afforded by my friend’s advice—specifically, the better odds of success—will still remain. So too, is it the case with deterrence. While an individual will be best off where she is never the victim of offending, she will be better off where the odds of her victimisation are lower. To say, then, that an individual is “guaranteed to receive a benefit” from a particular punishment practice is simply to say that she is guaranteed the benefit of having her likelihood of victimisation reduced. This benefit will stand regardless of whether or not she goes on to actually be victimised.

Thus, while guaranteed benefit may bring a certainty of better odds, it does not bring a certainty that an individual will actually be better off. She may be victimised both with—and without—the punishment practice in place. All of this is to say that we should not immediately discount expected benefit on the basis that it lacks certainty of actual benefit in this same way.
Expected Benefit

Suppose, then, that we pursue this idea of merely requiring an expected—not guaranteed—benefit in order to show that an individual has good reason to agree to a punishment practice. Such an assessment would be made ex ante, and would require the individual to evaluate her odds of being subjected to a particular punishment at some point in the future. In this way, the main focus would be on whether or not she had greater prospects of benefit under the practice, or under the Status Quo. This assessment does involve a number of complications that need to be addressed. For one, we might think that the punishment of an individual directly correlates with that individual’s decision to commit a crime, but this is simply not the case.

First, the commission of a crime is not sufficient for the receipt of punishment because many crimes go unpunished. The actual number of crimes that result in the punishment of an offender is given by a combination of two values: the resolution rate\(^\text{165}\) (the number of reported offences for which an offender is identified and charged), and the conviction rate (the number of charges that result in a successful conviction). In 2005,\(^\text{166}\) only 43.3% of recorded offences in New Zealand were resolved,\(^\text{167}\) with 65% of these cases resulting in a successful prosecution.\(^\text{168}\) All of this means that only 28.2% of recorded offences resulted in a successful conviction, and that 71.8% of crimes went entirely unpunished.\(^\text{169}\)

---

\(^{165}\) Referred to in some jurisdictions as the ‘clearance rate’.

\(^{166}\) The latest year for which resolution rate statistics are available in New Zealand. More up-to-date figures will be provided in the final version of this thesis.


\(^{169}\) It is difficult to provide similar statistics for the United States. While the Federal Bureau of Investigation records resolution rates on a per-offence basis (see The Federal Bureau of Investigation, “Offences Cleared” <www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/clearances>), conviction rates are instead provided by the U.S. Department of Justice on a *per-defendant* basis (see U.S. Department of Justice
Second, the commission of a crime is not necessary for the receipt of punishment, as gross miscarriages of justice can happen.\textsuperscript{170} Unfortunately, it is difficult to establish a specific rate at which they occur. This is primarily due to the fact that many miscarriages may go undiscovered or unproven. In 2006, Supreme Court Justice Antonin Scalia claimed that the false conviction rate for murder was somewhere around 0.027%\textsuperscript{171} A more detailed calculation by University of Michigan Law School Professor Samuel Gross instead conservatively places this value around the 4.1% mark—meaning that more than one in twenty-five individuals who are punished for murder are done so wrongfully.\textsuperscript{172}

Suppose, however, that a perfect justice system was possible—that all (and only) offenders were punished for all crimes, and that the punishment of an individual directly correlated with that individual’s decision to commit a crime. Even in this scenario, concerns may still arise. These will stem largely from the fact that an assessment of expected benefit will apparently require us to make reliable predictions about our future behaviour. Put another way, it will necessitate us estimating the precise odds of whether we will (or will not) decide to commit a certain crime at some indeterminate point in the future.

There are two central problems with this approach. The first of these is largely an epistemic one: gaining any sort of knowledge (let alone accurate knowledge) of what our future

\textsuperscript{170} To take one example, the Innocence Project has successfully used DNA evidence to exonerate 330 individuals post-conviction. See The Innocence Project, “The Cases: DNA Exoneree Profiles” <www.innocenceproject.org/cases-false-imprisonment >.


choices will be is notoriously challenging. Psychologies and circumstances change, so it is incredibly difficult—if not entirely impossible—for any individual to predict with certainty what she will do in the future. The second problem is instead a logical one. Derek Baker notes that “you cannot deliberate about whether to perform an action if you already know that you will—at least not rationally.”\textsuperscript{173} Given this:

…it would never make sense, then, for an omniscient being to reason about what to do. But an agent is a creature that reasons about what to do—which means that “omniscient agent,” must be an oxymoron.

It is this line of reasoning which leads Baker to his titular claim that “deliberators must be imperfect.” Stated in premised form, his argument appears as follows:\textsuperscript{174}

1. Necessarily, if an agent knows that she will $A$ and the agent is perfectly rational, then the agent cannot deliberate about whether to $A$.
2. Necessarily, for any perfectly rational agent, there is some possible $A$ the agent can deliberate about performing.
3. Necessarily, if an agent is omniscient, then for every possible $A$ the agent knows whether she will $A$.

C. Perfectly rational omniscient agents are impossible.

\textsuperscript{174} Baker, p. 3.
What this means, for our purposes, is that it is not only *epistemically challenging* for a rational agent to have concrete knowledge of his future choices, but also *logically impossible*. A rational agent who is fully cognisant of their future decisions will not, by definition, be capable of rational choice in any matter.

Fortunately, there is a way for us to modify our consideration of ‘expected benefit’ to avoid these problems—specifically, by showing that an agent can still reason intelligibly about expected benefits *without* requiring his omniscience. This eliminates both (1) the nearly impossible requirement of full and concrete information about future choices, and (2) any risk of entailing a logical contradiction. It is *not* contradictory for an imperfect deliberator to both make a prediction about a future action $A$ and still rationally deliberate about $A$.

So, if we are to assess our expected benefit *without* omniscient knowledge of our future choices, what might we look to instead? I suggest that it would be sufficient to merely look at our intentions as they currently stand, and as they would stand if we decided to agree to the punishment practice in question. While examining my current intentions, I would also need to be mindful of several other factors, including (1) the difficulty of following through with my intentions (including the likelihood of succumbing to temptation); and (2) the likelihood of my intentions changing at some point in the future. For anything less than an omniscient agent, these factors will be assessed within some margin of error. We cannot know with absolute certainty how fickle our dispositions will be. Thus, some uncertainty will inevitably creep in when we attempt to assess whether or not we can expect to benefit under the punishment practice in question.

But this uncertainty need not disqualify ‘expected benefit’ as a useful tool of establishing when we have good reason to agree to something. For one, the extent of the uncertainty can be
ameliorated by choosing the right time at which to make the assessment of expected benefit. The earlier we make a prediction (that is, the more temporal distance between the prediction, and the act predicted), the more unreliable the prediction becomes. This remains true when assessing the extent to which we expect to benefit from something. We might initially expect to benefit from a particular course of action, but no longer expect this after further information comes to light. Alternatively, we may now find ourselves expecting to benefit from something which, at some earlier point, seemed as though it would bring us no benefit. In the next chapter, I will consider precisely when an assessment of expected benefit should occur. I will propose several alternatives, before signalling my support for this assessment occurring at the most temporally relevant point—that is the last moment before which an individual decides to commit a crime.

Further, the uncertainty inherent in our ‘expected benefit’ assessment may actually incentivise more individuals to give their assent to punishment practices. Since no future outcome can be known with full certainty (including the outcome in which an individual is actually punished), no potential offender can ever take into account the full detriment of receiving punishment. Put another way, even the most likely of criminals will need to accept the fact that there is some chance that she will not—for any number of reasons—actually offend. So long as this is the case (that is, so long as it is the case that there is sufficient uncertainty such that the probability of her offending is a value less than 1) then the full disutility of the punishment will be discounted to at least some extent. This will increase the likelihood (even if only marginally) of an individual—particularly a potential offender—endorsing the implementation of a particular punishment practice.

As noted above, when considering what we expect ourselves to do, we must not focus solely on our current intentions, but also on what our intentions would become post-punishment
practice implementation. This is particularly important, as a punishment practice with a truly effective measure of deterrence will influence our intentions regarding the relevant offence. However much (or little) we currently intend to commit a crime, this intention will be further reduced by the existence of any effective deterrence. As such, there is an important reflexive relationship at play here. The kind and extent of punishment being proposed for an offence may very well have some impact on the likelihood that an individual will commit that offence. In this way, a potential offender might choose to tie herself to the mast in Ulyssean fashion, endorsing a punishment practice so draconian that it drastically reduces her chances of ever committing the offence in question, thus increasing her odds of experiencing the outcome of Beneficial Punishment from the practice in question.

**Having Good Reason to Agree to Punishment**

In this chapter I have outlined two potential standards for when we might have good reason to agree to a punishment practice. The first of these was guaranteed benefit. This standard requires us to show that an individual is guaranteed to fare better under a punishment practice than under the status quo—even in the worst-case scenario. Put another way, it requires showing that the utility gained from the deterrence of the crime is, for the individual, sufficient to outweigh any disutilities she suffers, including the disutility of $U_{Sanction}$ if she is punished. I claimed that if anything can provide us with a good reason to agree to something (and thus ground the moral permissibility of punishment) it seems clear that guaranteed benefit can.

Having established the sufficiency of guaranteed benefit, I then moved on to consider the possibility that a lower standard of expected benefit might also suffice. I argued that—despite
lacking the kind of concrete assurances provided by guaranteed benefit—expected benefit might also (with certain suitable additions in the following chapters) be a workable standard for an individual having good reason to agree to something.

We must now consider which of these thresholds to adopt for the remainder of our discussion. Essentially, the contractarian has two options: to either adopt the higher threshold of guaranteed benefit, or the lower threshold of expected benefit. Guaranteed benefit has the advantage of being much cleaner and simpler in application, though will provide us with far less in the way of implementable punishment practices. Expected benefit, on the other hand, will garner agreement for a wider range of punishment practices, though will require a more careful and complex analysis.

Since expected benefit only deals with what is likely to happen, the contractarian will have to establish precisely what kinds of risks it is rational for an individual to take. So far, however, our discussions of expected benefit have assumed that such risk-calculation will simply be calculated on the basis of expected utility—and this seems to have worked well. If an agent is presented with two courses of action $A$ and $B$, and calculates that she can expect a higher utility from choosing $A$, then she will clearly have good reason to choose $A$ (even if, at some future point, $A$ turns out to provide her with less utility than the alternative). For the purposes of this thesis, then, I will adopt expected benefit as the threshold for when an individual has good reason to agree to something.

I began this chapter by claiming that it will be morally permissible to punish an individual $P$ under a particular punishment practice $X$ so long as, at some prior time, $P$ simultaneously (i) had good reason to agree to $X$; and (ii) gave some kind of agreement to $X$. I have now argued that an individual will have good reason to agree to a punishment practice
where she can expect to benefit under that practice. But are there any actual punishment practices that would satisfy this requirement? It might turn out that there is little—if anything—that the contractarian theory of punishment justifies.

Clearly, there are conceivable punishment practices that would meet the standard outlined in the formula above. One of the most-straight forward examples is a case of punishment for murder. The valuation of our own life is perhaps one of the strongest values that many individuals hold. Further, its valuation is a prerequisite for many other values. We cannot value seeing the Astros win the World Series, or completing the next great American novel without necessarily also placing some value on our continued existence. Because of this, individuals are likely to tolerate large detriments in return for the overwhelming benefit of continued life. Suppose that there was a society in which individuals were exposed to a more than negligible chance of being murdered before reaching adulthood—say 10%. Imagine, further, that empirical evidence was obtained showing that the implementation of a mandatory ten-year prison sentence for murder would reduce the probability of being murdered to effectively 0%. In this case, it seems that all individuals would receive an overwhelming benefit in implementing this particular punishment practice. Even those who experienced Detrimental Punishment (i.e., were subsequently incarcerated for a decade) would still benefit from the practice—being better off alive in prison, than having never made it to adulthood in the first place. In other words, your punishment would be justified because the benefits of the deterrent scheme that enabled you to survive to adulthood:

---

175 Assuming that certain standards of treatment in prison are maintained. There are undeniably conceivable cases of incarceration that may be worse than death.
have been great enough to you, throughout your life, that they overwhelm even the disvalue you are presently experiencing from your ten-year sentence. Your life is still better than it would have been in the absence of that sentencing provision, despite the fact that it has resulted in your incarceration. The benefit, in short, is not just to society generally: It is one that attaches to each particular offender and supplies each with a ground for consenting to the deterrent scheme under which he is to be punished.

But what will the contractarian test say of those who can expect to be punished with a more severe penalty—say, those offenders who receive the death penalty or life in prison? Can such an individual possibly regard himself as better off than he would have been under the Status Quo? It is certainly possible. If a world without social cooperation—that is, the state of nature—is as brutal as Hobbes describes it, then “it is possible that a person receiving a very severe sentence like death or life in prison without parole would regard himself as benefitted as compared with his life in the absence of such penalties.”\(^{177}\) However, whether this is so would depend on a number of factors, including the marginal deterrent benefit of these serious punishments relative to more moderate penalties, and when in her life the offender receives the penalty.\(^{178}\) An older person who had many years to reap the deterrence benefits of the punishment practice would be in a very different position to a young offender “who now could reap no further benefits from such rules if put to death or imprisoned for the rest of his life.”\(^{179}\) The test of expected benefit is therefore highly context-dependent, and will necessarily require a case-by-case analysis. I will turn to do this in Chapter 5, demonstrating what kinds of punishment practices our theory would prescribe for the crimes of murder, grievous bodily harm, property theft, and digital piracy.

\(^{177}\) Finkelstein (2005), p. 216.
\(^{178}\) Finkelstein (2005), p. 216.
\(^{179}\) Finkelstein (2005), p. 216.
Having outlined in what circumstances an individual will have good reason to agree to a punishment practice, we can now turn to consider a much more troubling question: how does an individual actually give their agreement? This question will be the central focus of the next chapter.
Chapter 4: How We Agree to Punishment

In Chapter 2 I broadly outlined the contractarian theory of punishment, arguing that it will be morally permissible to punish an individual $P$ under a particular punishment practice $X$ so long as, at some prior time, $P$ simultaneously:

(i) Had good reason to agree to $X$; and
(ii) Gave some kind of agreement to $X$.

In the previous chapter I went on to analyse the first of these criteria—considering in what circumstances (if any) an individual will have good reason to agree to a particular punishment practice. I began with the claim that an individual will, at the very least, have good reason to agree to a punishment practice where she is guaranteed to benefit under that practice—that is, where she will fare better in a world with the practice, than without it, even if the worst-case scenario occurs (that is, the scenario in which she herself is subjected to the punishment). I then moved on to argue that a lower standard will also suffice—namely, one based upon the expectation of benefit. On this approach, an individual will have had good reason to agree to a punishment practice if, at some prior time, she could have expected to benefit under that practice when compared with the status quo. I claimed that in order to establish this benefit, we must carefully balance the practice’s utility in the form of $U_{\text{Deterrence}}$, against its various disutilities in the form of $U_{\text{Collateral}}$, $U_{\text{Restraint}}$, and $U_{\text{Sanction}}$.

In this chapter, I will turn to focus on the second of the above criteria—that is, the condition that an individual must previously have given some kind of agreement to the
punishment practice in question. This may seem like a straight-forward requirement, but it is fraught with complications. For one, there is some ambiguity about precisely what will count as “agreement.” Must it be written or verbal, or can it be implied by other less overt actions? Do we even require action for agreement? Might inaction be sufficient in some circumstances? In what follows, I will consider four potential kinds of agreement: express, implied, tacit, and hypothetical. I will argue that when it comes to a practicable theory of punishment, the first three of these are found wanting for a variety of different reasons. In light of this, I will contend that our contractarian theory is best underpinned by a foundation of hypothetical—not actual—agreement. In Chapter 5 I will attempt to allay some of the misgivings that commonly surround an appeal to this kind of agreement.

**Express Agreement**

The simplest form of agreement is that which is given expressly. Express agreement occurs when our words—whether written or verbal—clearly convey our intention to be party to a particular agreement. We make an express agreement when we make a promise, swear an oath, sign a treaty, or shake on a deal. If, for example, Alison asks Belinda if she can borrow her car for the afternoon, and Belinda replies with an incontrovertible “yes,” then Belinda will have given her express agreement for Alison to use her car.

It is important to ensure that we do not confuse an express agreement with a legal contract however, as the latter entails standards that are much higher. For one, a contract requires that each party provides consideration—that is, an actual or promised detriment to themselves, or an actual or promised benefit to the other party. If Charles asks David to mow his lawns, and
David agrees, he is *not* contractually bound to make good on this promise. This is because Charles has given no consideration. He has not—and will not—suffer any detriment, nor provide any benefit to David in return for his services. As such, David is not legally required to keep up his end of the agreement. If, however, Charles provides some sort of consideration—say, the promise of a cold beer in payment—then David will be bound.\textsuperscript{180} No such standard necessarily exists in the case of express agreement. While Charles and David’s deal may fail to meet the requirements of a legally binding contract, it may very well be sufficient for a morally binding agreement. This will, of course, depend on how the requirements for such an agreement are further fleshed out.

Suppose that we rely upon express agreement in the context of punishment. What might such a system look like? One possibility is that before implementing a particular punishment practice, the state would furnish all of its citizens with a written copy of a ‘punishment contract’. This contract would outline the exact parameters of the proposed punishment practice, including (1) the particular crime to be punished, and (2) the particular kind and severity of punishment that the crime will attract. Now, typically, the punishment meted out by a justice system will be modified depending on a number of relevant aggravating factors (such as the violence or cruelty involved, the vulnerability of the victim, any premeditation, and previous offending) and mitigating factors (such as the age, diminished intellectual capacity, remorse, and character of the offender).\textsuperscript{181} For this reason, the punishment contract would likely not include *specific* punishments, but rather outline penalties in terms of minimum and maximum sentences: for example, “a minimum term of imprisonment of two years—but no longer than five years—for committing crime C.”

\textsuperscript{180} The quality of this consideration is not considered—leading Lord Denning to famously note that a mere peppercorn would suffice for consideration (see *Chappell & Co Ltd v Nestle Co Ltd* [1960] AC 87).

\textsuperscript{181} See, for example, the Sentencing Act 2002 (New Zealand), s. 9.
An individual who freely signed this contract could be treated as having incontrovertibly agreed to its terms. The individual would then go on to enjoy the benefits of the contract (namely, its deterrent effects) while at the same acknowledging that she will be subject to the punitive measures outlined within the contract if she goes on to commit the offence in question. An instance of agreement like this would clearly satisfy the second criterion of the contractarian theory of punishment.

There is a problem, however: namely, that requiring this kind of agreement would be an incredibly onerous burden for any punishment system. To publish, disseminate, and retrieve a punishment contract from every citizen for every punishment practice would be an almost impossibly gargantuan task for any government to undertake. Now, it is true that our intention here is to formulate a normative theory—that is, a theory that focuses on how our punishment system should look. And it is also true that, in light of this, we are entitled to bite the bullet and hold that this onerously high standard of express agreement is required, and that a punishment practice will only be morally permissible where agreement is present in this form. At the same time, however, we would hope that our theory is partially descriptive—identifying at least some punishment practices that already exist out in the world as being morally permissible. But how often are express punishment agreements made between the individual and the state? The answer, it seems, is “rarely, if ever.” Indeed, A. John Simmons argues that “the paucity of express consentors is painfully apparent,” and that “most of us have never been faced with a situation where express consent to a government’s authority was even appropriate, let alone actually performed such an act.” It would be contentious to claim that no current punishment practice is morally permissible—but if we maintain that express agreement is the necessary

183 Simmons, p. 79.
standard for the moral permissibility of punishment, then it seems that this is exactly the conclusion that is drawn by our contractarian theory. In light of this, it is worth exploring whether some lower standard of agreement might instead be sufficient. If we can show that it is, we will leave the door open for some current punishment practices to be deemed morally permissible.

**Implied Agreement**

Instead of restricting ourselves to merely written or verbal signs of agreement, we might also look to a wider range of behaviours that suggest the consent of an individual. This is known as *implied* agreement. Implied agreement occurs where an individual fails to give express agreement, but otherwise acts so as to give a clear indication of her intention to be bound by the agreement. Suppose, for example, that when Alison asks Belinda to borrow her car, Belinda does not reply with a verbal answer, but instead hands over her car keys to Alison with a smile. Such an action could be taken as a clear sign of implied agreement. While Belinda does not explicitly say “yes,” her actions are sufficient to imply that she agrees to allow Alison to borrow her car.

This kind of agreement may prove very useful in the context of punishment, as it does not require a government having to go so far as ensuring the written or verbal agreement of every individual. Instead, it can observe and construe certain actions of its citizens as being clear signs of agreement. But what kinds of behaviour might count as implied agreement to a particular punishment practice? The most promising possibility is to hold that voting or otherwise participating in the democratic process is a sufficient demonstration of an individual’s
agreement.\textsuperscript{184} Suppose that a certain candidate campaigns strongly on introducing a particular punishment practice: the death penalty as a response to acts of treason, for example. We might then assume that any individual who votes for this candidate will have given her implied agreement to that punishment practice—despite the fact that she will not have expressly given her written or verbal agreement to its implementation.

There are, however, several problems with this approach. First, many justice systems allow for the punishment of individuals who are not yet of voting age.\textsuperscript{185} But since these youths are unable to vote—thus leaving them unable to register their implied agreement (or disagreement) to any kind of punishment practice—then it will morally impermissible to sanction them under any punishment practice. This, however, could easily be remedied. If implied agreement in the form of voting turned out to be the appropriate foundation for our theory, then it would simply be the case that those who are too young to vote could not be punished. This does not seem a highly controversial suggestion—particularly keeping in mind the fact that alternatives to punishment (such as rehabilitation) would still be open as permissible responses to young offenders. Indeed, the moral claim being made in regards to youths would be similar to that of “no taxation without representation,” but stated as “no punishment without representation” instead.

Nevertheless, there is a second, more fundamental, concern with using the act of voting as a ground for implied agreement. This problem arises from the fact that those running for political office rarely campaign on punishment in a way that is sufficiently comprehensive or detailed to form the basis of an acceptable punishment agreement. Candidates largely fail to


\textsuperscript{185} In New Zealand, children as young as 10 years of age may be convicted of a crime (see Crimes Act 1961 (New Zealand), ss. 21 and 22), while only those 18 or older are eligible to vote.
propose specific punishments for specific crimes, and instead tend to couch their punishment policies in attractive aphorisms such as “tough on crime” or “promoting victims’ rights” with little further elaboration on what those positions actually entail. They will not usually speak in terms of specific punishments and types of sentences. On the rare occasions that such specifics are given, they will usually only cover a small number of offences—omitting details on how a vast number of other crimes are to be dealt with. For this reason, implied agreement in the form of voting will only be available for a very limited number of punishment practices.

But suppose that a candidate was to campaign on a platform that outlined a comprehensive and detailed range of punishment practices. Even if this was to happen, it is still not clear that voting for that candidate would be sufficient to indicate implied agreement with her proposed punishment practices. This is because candidates typically campaign on a range of varied issues. A voter might detest a candidate’s policies relating to punishment, but nevertheless find the candidate’s positions on other issues—like health, education, and immigration—attractive enough to deserve a vote. In this case, the act of voting for the candidate would not be sufficient to demonstrate implied agreement with the punishment policies of the candidate in question.

In light of these observations, then, implied agreement in the form of voting appears to largely fail as a solid basis for the contractarian theory of agreement. Candidates’ punishment policies are rarely—if ever—detailed or comprehensive enough to form the foundation of a punishment agreement. Even where they are, the act of simply voting for a particular candidate is not sufficient for expressing implied agreement. For this reason, we need to consider further alternative sources of agreement.
Unanimity and Tacit Agreement

So far, we have seen how both express and implied agreement fail to give us a satisfactory way of identifying when an individual gives his or her agreement to a punishment practice. Express agreement sets the bar too high for agreement, while implied consent is neither necessary nor sufficient for agreement. At this stage, it is worthwhile discussing a further concern that plagues both of these forms of agreement—namely, that even if these forms of consent were appropriate indicators of agreement, the breadth of obtained agreement would be insufficient to provide a moral basis for a wider punishment practice.

In order to illuminate this issue, it is necessary to remind ourselves that there are two very different questions to examine when considering the moral permissibility of punishment:

1. **Is a particular instance of punishment** morally permissible?; and
2. **Is a wider punishment practice** morally permissible?

When considering (1), we are asking whether or not it is morally permissible to sanction person $P$ for offence $O$ with punishment $X$. We might ask, for example, “is it morally permissible to sanction $P_1$ for murder by executing him?” or “is it morally permissible to sanction $P_2$ for stealing a loaf of bread by fining him one hundred dollars?”

When considering (2), we are instead making inquiries into whether or not it is morally permissible to implement a practice that will sanction *every* individual who commits offence $O$ with punishment $P$. We might ask, for example, “is it morally permissible to sanction any
individual who commits murder by executing him/her?” or “is it morally permissible to sanction any individual who steals a loaf of bread by fining him/her one hundred dollars?

While we might be able to morally justify a single instance of punishment as it applies to one solitary individual, it will be much harder to morally justify a practice that will punish *all* individuals in this same way. So far, we have dedicated much of our time to discussing when a particular instance of punishment will be justified. The conclusion we came to is that with which I started this chapter—that it will be morally permissible to punish an individual under a particular punishment practice so long as, at some prior time (i) the individual had good reason to agree to that punishment practice; and (ii) the individual gave some kind of agreement to that punishment practice. This is our answer to (1) above.

In answering (2), we merely need to generalise this answer to all members of society, and claim that *a wider punishment practice will be morally permissible where the above test is satisfied for all individuals within the community over which the practice applies*. Focussing specifically on (ii) for a moment, this means that a morally permissible punishment practice will require agreement from *everyone* to whom it applies.

This is a high threshold, so it is worthwhile considering why we might need such a burdensome standard. For one, K.L. Avio notes that “[t]he basic rights of individuals within society are determined at the constitutional level of discourse” and that “[u]nanimity is required for determining these rights since veto power is the condition imposed by individuals for entering into civilised society.”  

The implementation of any punishment practice will potentially involve certain encroachments on these basic rights. Punishment—as defined in Chapter 1—necessarily involves harm, and this harm may come by depriving individuals of their

---

property (as in the case of fines), their liberty (as in the case of incarceration), and even their lives (as in the case of capital punishment). The fact that punishment may so readily interfere with individuals’ rights to these things seems to suggest that—on Avio’s approach—a morally permissible punishment practice will require unanimous agreement from everyone who, at any point in the future, might be subjected to its sanctions.

If we find this argument unpersuasive, then there is another, more practical reason why we should seek unanimity: namely, that we do not want an ad-hoc system where some individuals have given agreement to a punishment practice (and are therefore subject to its sanctions) while others are not. In fact, it is difficult to imagine exactly what such a system would even look like. If an individual fails to agree to a punishment practice, then she should not be granted its benefits, lest we invite free-riders. But the deterrence benefits created by punishment practices come in the form of gross reductions in the number of crimes being committed. A successful punishment practice for the crime of vehicle theft will see an appreciable reduction in the rate of vehicle thefts across society. In this way, it seems that non-agreeing parties will be capable of benefitting from this reduction regardless of whether or not they agree to the deterrence-creating punishment practice. One way around this might be to withhold the benefit of deterrence from those who have not agreed to a practice—that is, to not punish crimes committed against those individuals who have not agreed to the practice, and to make this widely known to potential offenders. But this would be a bizarre system indeed.

So, while we want our punishment practices to be morally permissible, we also want those practices to have universal application. In order for a universal punishment practice to be morally permissible under the contractarian theory, it should have the consent of all individuals within its jurisdiction—that is, unanimous agreement. If we can obtain this, we ensure that each
particular instance of punishment under a practice is made morally permissible by the fact that the *very individual being punished* has previously given agreement to the wider punishment practice.

It is unlikely that either express or implied agreement would ever garner unanimous agreement. To have an entire population agree to a specific punishment practice—whether by their words, or actions—would be an enormous feat. Fortunately, there is a third mode of consent that is well-placed to ensure unanimity: namely, tacit agreement.

Whereas both express agreement and implied agreement require that we do *something* to register our consent, tacit agreement instead requires that we do *nothing*. Put another way, while express and implied agreement are based upon action, tacit agreement is instead based upon *inaction*. In cases of tacit agreement, an individual is assumed to have given her consent provided that she has not acted so as to clearly express otherwise. Consider the case in which a priest asks a congregation if anyone objects to the marriage that he is about to sanctify. In this scenario, the failure of the attendees to speak up amounts to their *tacit agreement* to the union. In this way, agreement is given “by remaining silent and inactive; it is not express, explicit, directly and indirectly expressed by action, but rather is expressed by the failure to do certain things.”

In order to operate effectively, tacit agreement requires the identification of some particular action or actions that can be performed to ‘opt-out’ of an agreement. In the case of the marriage example, opting-out might require something as simple as standing up and shouting “I object.” It is the absence of the exercise of this opting-out procedure that will lead us to assume that all present parties agree. What, then, might the particular opt-out method be when it comes to punishment practices? The simplest approach would be to require something similar to the marriage scenario—that is, a clear and unequivocal expression of an individual’s disapproval of

---

187 Simmons, p. 80.
the punishment practice. So, if a citizen did not wish to be bound by a punishment agreement, he could register this desire by, say, informing his government representative. Once he had done this, his tacit agreement could no longer be assumed, and it would be morally impermissible to punish him under that punishment practice.

There are problems with allowing such a simple opt-out, however. Specifically, this version of tacit agreement begins to fall prey to some of the same problems we witnessed with express and implied agreement. Individuals will be able to withdraw from punishment agreements with relative ease—but it is less clear how we can then ensure that these non-agreeing parties do not receive the *benefits* of the practice. As noted above, deterrence occurs wholesale, and it is difficult to imagine how the system might be modified to as to avoid non-agreeing parties becoming free-riders.

A better approach might be to adopt the strategy of Hobbes, \(^{188}\) Locke, \(^{189}\) and Rousseau \(^{190}\) and establish tacit agreement on the basis of *residence*. On this view, an individual would signify her disagreement with a punishment practice by choosing to remove herself from the jurisdiction in which that practice had been implemented. If an individual instead chooses to *remain* in a particular geographical area, then we can assume that she tacitly agrees to those punishment practices that are in effect there. This approach has the benefit of providing us with an incredibly economical answer to the second criterion of the contractarian theory of punishment (that is, the requirement that the individual previously gave some kind of agreement to that punishment practice). Establishing whether or not agreement has been given will be relatively easy—we merely need to look at who remains resident within a particular jurisdiction, and can assume that all of those individuals have given their agreement to whatever punishment practices are in effect.

\(^{188}\) Hobbes, Part II, Chapter 20, p. 253.

\(^{189}\) Locke, Chapter VII, §119, p. 64.

over that jurisdiction. Tacit agreement by residence has the further advantage of guaranteeing unanimity within a jurisdiction—for, if disagreement is registered by non-residency, then all who remain resident within the jurisdiction of the punishment practice are (by definition) counted as having given their agreement.

But while tacit agreement on the basis of residence works well in theory, its practical application is multiply problematic. Simmons notes that inaction can only be the basis of tacit agreement where five distinct criteria are met:\textsuperscript{191}

1. The situation must be such that it is perfectly clear that consent is appropriate and that the individual is aware of this. This includes the requirement that the potential consentor be awake and aware of what is happening.

2. There must be a definite period of reasonable duration when objections or expressions of dissent are invited or clearly appropriate, and the acceptable means of expressing this dissent must be understood by or made known to the potential consentor.

3. The point at which expressions of dissent are no longer acceptable must be obvious or made clear in some way to the potential consentor.

4. The means acceptable for indicating dissent must be reasonable and reasonably easily performed.

5. The consequences of dissent cannot be extremely detrimental to the potential consentor.

\textsuperscript{191} Simmons, pp. 80-81.
These conditions place certain constraints on how our system of tacit consent by residence might operate. We could not, for example, implement a new punishment policy in the middle of the night and claim that sleeping individuals were giving their tacit agreement simply because they were not objecting. This would be in violation of condition (1). Further, we could not take an individual’s residence immediately after the implementation of a punishment practice as a sign of tacit agreement. This would be in violation of condition (2). Changing residence takes time, so some ‘reasonable duration’ would be necessary.

Nevertheless, it seems that our system could be suitably (and relatively easily) structured in order to satisfy at least the first three of these conditions. The real concern is instead with conditions (4) and (5). Given the real-world operation of states and nations, it is not clear that tacit agreement by residence would ever be capable of meeting these conditions. Gone are the days when an individual could easily opt-out of a society and depart for the wilderness or the frontier. These days, becoming a non-resident of one jurisdiction necessarily requires becoming resident of another—and this is no easy task. In many cases there are insurmountable financial or legal barriers to becoming a non-resident. In other cases an individual may be forbidden from leaving her own jurisdiction, or forbidden from entering another, depending on her past behaviour and the laws of the states involved. Even if such a move is possible, it may still be highly onerous in light of other considerations. Factors such as politics, culture, and religion all come in to play, not to mention familial and community ties. An individual may strongly disagree with a punishment practice, yet opting-out via non-residency may require other sacrifices that are simply too great. In fact, the insufficiency of tacit agreement by residence is very similar to the insufficiency of implied agreement by voting. While both actions (voting and residence) might coincide with consent to the punishment practice, they are not sufficient for it.
Just as an individual may have other reasons for voting for a candidate whose punishment policy she does not agree with, so too might an individual have other reasons for remaining resident in a society with a punishment practice she does not endorse. For this reason, it seems too quick to assume that an individual’s decision to remain resident in a jurisdiction under a particular punishment practice necessarily means that she has given her tacit consent to that practice. In light of this, we need to continue our search for some other kind of appropriate agreement.

**Hypothetical Agreement**

So far, I have argued that express, implied, and tacit agreement all fail as a satisfactory basis for a contractarian theory of punishment. Despite their differences, these three forms of consent have one thing in common—they all look to some historical action or inaction in order to show that actual agreement has been given. I therefore propose an alternative approach: that is, to argue that the unanimity required for the moral justification of a punishment practice arises from the hypothetical agreement of those to whom the practice applies.

The usage of ‘hypothetical’ agreements is common within the social contract tradition. Stephen Darwall notes that:

> “only rarely do contractarians or contractualists claim that moral standards depend on actual agreements. More frequently what they hold is that moral principles are those that would be rationally or reasonably chosen or agreed to under certain ideal, counterfactual conditions.”

When we say that someone *hypothetically* agrees to something, we mean, roughly, that she *would* have agreed to it had she been asked under certain suitable conditions. Imagine, for example, that I go out of town to attend a conference and therefore have to leave behind my prized (and housebound) British shorthair kitten—Fizzgig. I ask my neighbour to let herself into my home once a day to feed Fizzgig, but do not give her permission to do anything further—and especially not to let Fizzgig leave the house. Suppose, then, that at some point during my trip Fizzgig is struck down with a mysterious illness. Unable to contact me, my neighbour makes the executive decision to take Fizzgig to the vet where he is promptly cured of an ailment that otherwise would have taken his life. My neighbour did not have my permission to remove Fizzgig from my home—but she could claim that she was justified on the basis of my *hypothetical agreement*. Put another way, she could argue that she is exculpated by the fact that *if* I had been aware of Fizzgig’s situation, and *if* my neighbour had asked for my permission to take Fizzgig to the vet, I would most assuredly have said “yes.”

Let us take this approach, then, and apply it to the case of punishment. To say that an individual hypothetically agrees to a punishment practice is to say that she *would* agree to the practice if asked. Stated roughly, then, the test for hypothetical agreement in the context of punishment might look something like this:

**HA1:** An individual hypothetically agrees to a punishment practice if she would give her actual agreement to the implementation of the practice when asked.
It is important to note, however, that the kind of contractarian hypothetical agreement I am suggesting here differs significantly from that proposed by authors in the *contractualist* tradition. Contractualists typically argue that certain societal structures would be hypothetically agreed upon by individuals who are ignorant of their own personal position within that society. The paradigm case of this is the ‘veil of ignorance’ proposed by John Rawls in his contractualist theory of justice.¹⁹³ The contractarian approach instead involves the argument that individuals *fully cognisant of their contingent features* would still hypothetically agree to certain societal structures. Such an approach is followed by David Gauthier, who imposes no veil of ignorance and “does not deprive his hypothetical contractors of knowledge of who they are and where they find themselves.”¹⁹⁴ Unlike their contractualist counterparts, then, contractarian hypothetical contractors know enough about their circumstances to be able to benefit themselves.

But why is such a move necessary? Why must our hypothetical contractors be given this knowledge? As noted in Chapter 2, the very basis of the contractarian theory is the claim that individuals are rational agents seeking to further their own interests (the objects of which might sometimes be the welfare of others). According to this theory, an individual will only have reason agree to a bargain that *actually* promotes her interests. A hypothetical agreement made behind a veil of ignorance will thus have little motivational force for an individual who—now aware of her own personal circumstances—calculates that she would have been better off not making the agreement. An individual will only be able to establish what is in her best interests (and thus what she should agree to) where she is fully aware of her own personal circumstances. Thus, in order to justify the implementation of a punishment practice, we need to provide each individual with a compelling answer to the question “what’s in it for me?” And this can only be

---

¹⁹³ See Rawls (1971).
done by imbuing hypothetical contractors with all of the characteristics that might be relevant to
the decision-making processes of their real-world counterparts.

But does the absence of this veil make much of a difference when it comes to the kinds of
punishment practices that will be agreed upon? I contend that it does. Behind a veil of ignorance,
most individuals might agree in principle to an institution of punishment and giving freedom to
legislators to implement punishment practices as they see fit. Without the veil, however, things
will be different. Certain individuals—for example, those from ethnic or socio-economic groups
that find themselves routinely targeted and disproportionately punished—will have a strong
interest in the specific structure of the punishment system. These individuals are instead only
likely to agree to an institution of punishment in which legislators are permitted to implement
punishment practices within certain parameters. Our test, then, can be modified to reflect the
extra personal knowledge that contractarian hypothetical contractors are assumed to possess:

HA2: An individual hypothetically agrees to a punishment practice if she
would—in light of her own contingent characteristics and circumstances—
give her actual agreement to the implementation of the practice when
asked.

It is worthwhile being explicit about precisely what these contingent ‘characteristics’ and
‘circumstances’ include. By including these elements, I wish to ensure that our hypothetical
contractors are not abstracted individuals, but fully realised ‘thick’ persons—doppelgangers of
their real-world counterparts in many important ways. At the very least, we want our
hypothetical persons to maintain important features such as their age, gender, ethnicity, religion,
and socio-economic position. The set of facts that comprises an individual’s ‘characteristics’ goes much wider than this, however, and also includes her full set of goals, values, and motivations. In fact, it includes every internal personal feature that might have some bearing on the calculation of whether or not she can expect to benefit under a proposed punishment practice. The set of facts that comprise the ‘circumstances’ of the individual is sketched in much the same way, and includes any external feature that may influence the same expected benefit calculation. Chief among these will be the social context in which the individual finds herself, including the likelihood that she may be unfairly targeted upon the basis of one of her characteristics.

By including the contingent characteristics and circumstances of the individual, I am assuming a conception of a person as—primarily—a collection of values and interests. The hypothetical agent I am here fleshing out thus does justice to this conception of a person, capturing what is most important to her rational decision-making facilities. What I do not wish to do, however, is ‘thicken’ my hypothetical agent to the point of reflecting the exact actual will of the individual. Why not? Because in order to truly establish what an individual would hypothetically agree to, we may need to abstract away from the agent’s actual will in certain subtle ways. For one, we want to focus on what these thick individuals would agree to when given full knowledge of the necessary facts. As noted in Chapter 3, the personal benefit of a punishment practice for an individual rests on the balancing of a number of relevant utilities and disutilities. Unfortunately, many of us can be greatly mistaken in our estimations of these. For one, surveys in the United States, Canada, Australia, and Britain have shown that the general public “has an inaccurate and negative view of crime statistics, including the [mistaken] general beliefs that the crime rate is rising dramatically and that a high percentage of crime involves
violence.” A similar trend is evident in New Zealand. While only 13% of all crimes committed are violent crimes, two-thirds of individuals surveyed estimated this proportion as being 50% or higher. Similar errors occur when judging the rates of certain crimes. To take one example, roughly 7% of households will be burgled in an average year in New Zealand. When surveyed, however, one-third of the population estimated this rate as being between 20% and 49%, while another third estimated it as being 50% or higher. Mistaken beliefs about the proportion and rates of certain crimes may lead us to make unwise bargains when it comes to deterring these kinds of offences—perhaps agreeing to far harsher penalties than necessary.

We can also be mistaken in other ways. For one, contracting agents might misjudge the costs of certain punishment practices. In New Zealand, the cost of keeping a person in prison for one year is around NZ$55,000. But when given four options to choose from (NZ$15,000, NZ$35,000, NZ$55,000, and NZ$75,000), only 41% of individuals correctly estimated this cost. Around 29% underestimated the cost, while 31% overestimated the cost at NZ$75,000. We may also be unaware of the true cost of certain crimes—both in circumstances where we are the direct victims of an offence, or where we end up paying for criminal activity in other ways (through increased insurance premiums, for example).

Another area for potential error is in estimating the effectiveness of a punishment practice—that is, the likelihood that a practice will deter (and therefore reduce) the rate of offending. Individuals may overestimate the effectiveness of particularly harsh penalties, or underestimate the effectiveness of less severe approaches.

196 Paulin et. al., p. 11.
197 Paulin et. al., p. 12.
198 Paulin et. al., p. 13.
Our test, then, needs to be revised so as to cover all of the potential ways in which a contracting agent might be misinformed about facts that will have a bearing on their expected benefit calculations. We can do this by amending it in the following way:

HA3: An individual hypothetically agrees to a punishment practice if she would—in light of her own contingent characteristics and circumstances, and with full knowledge of the relevant utilities and disutilities of the practice—give her actual agreement to the implementation of that practice when asked.

But, even armed with all of the facts, individuals still sometimes make decisions that fail to take into account their actual interests. They fail to be rational. This irrationality may be as a result of emotional instability, mental defect, or myriad other causes—but whatever the origin, it is something that we need to erase from the process of hypothetical agreement. Given this, our test needs one final tweak:

HA4: An individual hypothetically agrees to a punishment practice if she would—in light of her own contingent characteristics and circumstances, and with full knowledge of the relevant utilities and disutilities of the practice—give her actual agreement to the implementation of the practice when asked, and while acting rationally.
There is a potential tension here, however. In one breath, our test claims that we should take into account all contingent characteristics of the person—including psychological facts. In the next breath, however, the test also claims that we should assume the hypothetical contractor to be acting rationally. But what if these two are inconsistent? What if our hypothetical contractor suffers from a deep, long-term emotional instability that causes her to consistently act against her own values and goals? In assessing what this individual would hypothetically do, it seems that we must sacrifice one of the requirements of our contractarian test—that is, we must either ignore one of her contingent characteristics (her long-term emotional instability) or allow her hypothetical self to act irrationally.

For starters, the rationality condition is one that we simply cannot remove from our test. As noted above, the very basis of the contractarian theory is the claim that individuals are rational agents. To allow irrationality into our calculus will be to undermine this very foundation. If something is going to give, then, it must be the claim that we are to take into account all contingent characteristics of the individual.

So on what principled reason, then, can we take into account most of an individual’s contingent characteristics, and ignore certain others? It is clear that we only really run into an issue when one of these contingent characteristics infringes on the rationality of the individual. So, in order to not cast our net too wide, it is perhaps prudent to restrict our exclusion of contingent characteristics to only these cases. That is, to cases where a characteristic directly compromises an individual’s ability to make a rational choice.

Can we make this move without undermining the importantly ‘thick’ individuals that are so crucial to our contractarian theory of punishment? I contend that we can. In order to see this, it is important to remind ourselves precisely why we held the thickness of individuals to be so
important in the first place. At bottom, I claimed that this was so our theory could provide an
individual with a compelling answer to the question “what’s in it for me?” If our hypothetical
assessment ignores the individual’s values, or attributes, or the social context within which they
reside, then it will be hard-pressed to provide a reason that will have real motivational power for
the contractor. The individual will always be able to reply with something like: “while it is nice
that an abstracted version of me would hypothetically agree to this, that does not take into
account certain factors that influence what I would actually agree to.” The more contingent
characteristics and circumstances that we allow the hypothetical contractor to retain, the better
chance our theory has of providing the real individual with good reason to now adhere to that
agreement.

But removing a contingent characteristic that merely infringes on an individual’s ability
to make rational decisions does not seem to jeopardise our theory in the same way. The
hypothetical contractor will still retain all of those characteristics and circumstances of the kind
that we would ordinarily hold to be relevant to the decision-making process. The hypothetical
contractor will still reach the very best decision for the individual, given all of these
characteristics and circumstances. In fact, the hypothetical contractor may be better placed to
make this decision than the actual individual, as she is unencumbered by the malady that would
otherwise impair her rational abilities. Given this, it seems excusable to exclude from the set of
contingent characteristics those characteristics which directly and adversely interfere with an
individual’s ability for rational decision-making.

Having resolved this tension, we can now put forth HA4 as our full and final test for
hypothetical agreement in the context of punishment. If this test is satisfied, then an individual
can be taken as agreeing to the punishment practice for the purposes of the contractarian test for the moral permissibility of punishment.

At this point there may be some consternation about the moves we have made. The driving principle behind the contractarian theory is the claim that voluntariness has special moral force. Now, however, we have shifted our focus away from the actual agreements that individuals make, to what they may hypothetically agree to in certain idealised circumstances. This move away from actual agreement may foster fears of a paternalistic system—one which foists agreements on individuals irrespective of their express desires. In the next chapter, I will attempt to allay these concerns. I will return to my three initial desiderata—the features which, I argued, made a contractarian theory of punishment so appealing in the first place—and show that a theory based on hypothetical agreement still satisfies each of these criteria.

**When We Agree to Punishment**

As it stands, our contractarian theory of punishment holds that it will be morally permissible to punish an individual $P$ under a particular punishment practice $X$ so long as, at some prior time, $P$ simultaneously:

(i) Had good reason to agree to $X$; and  
(ii) Gave some kind of agreement to $X$.

In the previous chapter I have argued that an expectation of benefit will be sufficient to satisfy (i), while in this chapter I argued that hypothetical agreement will be sufficient to satisfy (ii).
One question looms large, however: namely, at what point in time must this dual-criteria test be satisfied?

So far, I have ambiguously stated that the test must merely be satisfied “at some prior time.” Now, however, it is necessary to become more explicit about this requirement. This is important, because the answer as to whether or not the test is satisfied for an individual may be largely contingent on when the assessment is made. It is clear, for example, that an individual who previously could expect to benefit from a punishment practice may no longer do so once he has been apprehended for committing the crime in question.

As I see it, there are three potential approaches we can take to this temporal issue: First, we can merely hold that it will be sufficient for the test to be met at any prior time. So, if many years ago $P$ both (i) expected to benefit from $X$, and (ii) hypothetically agreed to $X$, then this will be sufficient to now render punishment of $P$ under $X$ morally permissible. This is because agreements are, by their very nature, irrevocable (more on this in Chapter 6). So, once made, an agreement remains in force in perpetuity, regardless of the desires of the parties. If circumstances are such to give rise to a binding agreement in the past, then that agreement will persist through to the present moment. Thus, the satisfaction of (i) and (ii) at any prior time will be sufficient to give rise to an enduring moral justification for punishment that stretches into the future.

This approach may have some unattractive ramifications, however. For one, we can conceive of individuals who may spend a lifetime with no expectation of any benefit under a punishment practice, and no grounds for hypothetically agreeing to that practice. Perhaps, however, their circumstances change for the briefest of moments—and in that moment, they suddenly do have an expectation of benefit and do hypothetically agree to the practice in question. This fleeting instant will be sufficient to give rise to a perpetually enforceable
agreement, even if—immediately after that brief moment of expected benefit—they continue to live a life in which they can expect no benefit whatsoever from the practice.

A second approach to the temporal issue is to stipulate a specific point in time at which individuals would be conceptually polled for having satisfied (i) and (ii). A prime time for such a survey would be when a punishment practice is actually implemented. Thus, it will be morally permissible to punish P under X only if—at the time of X’s implementation—it was true that P both (i) could expect to benefit under X, and (ii) hypothetically agreed to X. Of course, for any individual, there will be a number of punishment practices that are in place before his or her coming into existence. How are we to approach these cases? It seems that when considering existing punishment practices, the best approach would be to assess an individual’s satisfaction of (i) and (ii) as soon as they attain the age of majority—that is, at the first instant at which they might potentially be subjected to the sanctions of the practice.

This is far from a satisfactory manoeuvre, however. As stated earlier, I have—for the purposes of this argument—adopted a subjectivist theory of value. Given this, there may be some concern surrounding the fact that the values of an individual who has just attained the age of majority may differ greatly to the values she will hold later on in life. It may therefore seem inappropriate to allow that younger individual to bind his or her future self to something so serious as a punishment practice.

A third approach is to carry out our assessment of the contractarian test at the very last moment before which an individual decides to commit a crime addressed by the punishment practice—that is, at the last possible moment at which an individual might still expect to benefit
from the practice. One complication with this approach is that it seems to suggest that only those individuals who are punished are taken as giving their hypothetical agreement. Since the contractarian test is only assessed upon an individual either (1) deciding to commit a crime, or (2) being wrongfully accused of a crime, those who go unpunished will never be polled for hypothetical agreement. This may be an acceptable ramification, however. It is the harm against those who are punished that we are seeking to provide a moral justification for, so it might be argued that it is only these harmed individuals whose agreement we need to garner. This is not strictly true however. Even those who are not punished receive certain detriments—for example, in the forms of $U_{Restraint}$ and $U_{Collateral}$. Given this, we need to ensure that these individuals also agree to the punishment practice that bestows such detriments upon them. The simplest solution to this problem might simply be to say that the assessment for individuals who have not committed a crime, nor been falsely accused of a crime, should always be taken as being carried out at the present instance.

With these three options in mind, we can now narrow down our approach to say that it will be morally permissible to punish P under X if (i) P could expect to benefit from X, and (ii) P hypothetically agreed to X at:

1. Any prior time; or
2. The latter of (a) the time that X was implemented, or (b) the time that P reached the age of majority; or
3. At the last moment before which P decided to commit a crime—or an innocent individual was wrongfully accused of a crime—covered by X.

---

199 Our theory does allow for the fact that some individuals may be punished for crimes they did not commit, so an analogous point in time—say, the moment before an individual is wrongfully accused—should be used in those kinds of cases.
I present all three of these options as potential routes to the reader. Any of these will provide a working basis for my theory. I, however, tend to favour the latter. This option avoids the possibility of a fleeting transient moment of expected benefit signing an individual up to a punishment agreement that can—at all other points in her life—be expected to bring her an expected detriment. Likewise, this option also prevents the scenario in which an adult is bound to a punishment practice by the transient values of their younger (and more immature) self. Indeed, in making the contractarian assessment, the third approach focusses on *the most relevant point in an individual’s history*—that is, the point in time immediately prior to the individual gaining knowledge that they will soon become subject to the sanctions of the punishment practice. It is this prior individual whose values most closely align with those of the individual who will actually go on to be punished.
Chapter 5: The Contractarian Theory of Punishment in Practice

So far, our treatment of punishment has occurred at a fairly abstract level. It is important to remember, however, that the sole purpose of this discussion is to establish a test by which we can implement punishment practices that are both morally permissible and practicable. In this chapter, then, I will focus on ‘road-testing’ the contractarian theory I have outlined—demonstrating how it might approach punishment for the crimes of murder, grievous bodily harm, theft, and digital piracy.

Before doing this, however, it is helpful to consolidate the discussions of the three previous chapters into a single, definitive contractarian test for a morally permissible punishment practice. Once this is done, I will also address two potential concerns with the test as it stands, before then moving on to apply it to the real-world crimes mentioned above.

Consolidating the Contractarian Test

In Chapter 2, I argued that a contractarian theory of punishment would make something like the following claim:

It will be morally permissible to punish an individual $P$ under a particular punishment practice $X$ if, at some prior time, $P$ simultaneously:

(i) Had good reason to agree to $X$; and

(ii) Gave some kind of agreement to $X$. 
In Chapter 3 I claimed that the first of these criteria will at least be met where an individual could expect to benefit under a punishment practice. In light of this claim, we can swap out (i) for (i*) as follows:

It will be morally permissible to punish an individual $P$ under a particular punishment practice $X$ if, at some prior time, $P$ simultaneously:

(i*) Expected to benefit under $X$; and

(ii) Gave some kind of agreement to $X$.

Likewise, in Chapter 4 I noted that while express consent would satisfy the second criterion, this would raise the bar too high for a comprehensive and practicable punishment system to attain the status of moral permissibility. I therefore argued that (ii) will also be satisfied where an individual merely hypothetically agrees to the punishment practice in question. This hypothetical agreement will be taken to have occurred wherever an agent would give her actual agreement to the implementation of the practice (1) in light of her own characteristics and circumstances, and (2) with full information regarding the relevant utilities and disutilities of the practice, and (3) while acting rationally. In light of this claim, we can swap out (ii) for (ii*) as follows:

It will be morally permissible to punish an individual $P$ under a particular punishment practice $X$ if, at some prior time, $P$ simultaneously:

(i*) Expected to benefit under $X$; and

(ii*) Would—in light of her own contingent characteristics and circumstances, and with full knowledge of the relevant utilities and disutilities of the practice—have given her actual agreement to $X$ when asked, and while acting rationally.
Finally, at the conclusion of the last chapter, I specified the time at which this assessment should be made for those who find themselves sanctioned under $X$. If we incorporate this, our test becomes the following:

It will be morally permissible to punish an individual $P$ under a particular punishment practice $X$ if, at the last moment before which $P$ decided to commit a crime covered by $X$ (or an innocent individual was wrongfully accused of a crime covered by $X$), $P$ simultaneously:

(i*) Expected to benefit under $X$; and

(ii*) Would—in light of her own contingent characteristics and circumstances, and with full knowledge of the relevant utilities and disutilities of the practice—have given her actual agreement to $X$ when asked, and while acting rationally.

This represents the final version of the contractarian test for the moral permissibility of punishment. Nevertheless, this test only tells us when a specific instance of punishment is allowed. As noted in the previous chapter, the implementation of a wider punishment practice over a community will only be morally permissible where this requirement is met for all individuals within that community.

We could go about applying this test as it stands, but there are two concerns that I wish to address before doing this. The first relates to the potentially superfluous nature of agreement, given the specific requirements of the contractarian test above. The second concern instead relates to the high bar of the unanimity requirement just mentioned.
Has Agreement Become Superfluous?

A distinctive feature of the contractarian theory of punishment is that it places significant moral weight on whether or not an individual has given her agreement to the punishment in question. A contractor is asked to consider “what’s in it for me?” and then to give (or withhold) her agreement depending on the answer to that question.

There is a concern, however, that this important element of agreement may have been lost, given the moves we have made in the preceding chapters. We have stated that the agreement condition of the contractarian test will at least be satisfied by hypothetical agreement. This is taken to occur whenever an individual would “in light of her own characteristics and circumstances, and with full knowledge of the relevant utilities and disutilities of the practice—give her actual agreement to the implementation of the practice when asked, and while acting rationally.” It seems, however, that rational individuals who meet this requirement (that is, who are given full information about the relevant utilities and disutilities of a practice) will at least hypothetically agree to the implementation of those punishment practices under which they can expect to benefit. This means that the satisfaction of criterion (i) seems to also entail the satisfaction of criterion (ii).

Now, this is not to say that the second criterion of the contractarian test for punishment is unnecessary. In fact, there may be cases where (ii) is satisfied while (i) is not. The inverse is not true, however. In all cases where (i) is satisfied—that is, in all cases where an individual can expect to benefit under the proposed punishment practice—(ii) will also be satisfied (that is, the individual will hypothetically agree to the proposed punishment practice). But, in light of this observation, it becomes apparent that we can assess whether or not the contractarian test for the
moral permissibility of punishment is satisfied merely by asking the first question: could the individual previously expect to benefit under the punishment practice? If the answer to this question is “yes,” then both criteria of the contractarian test for punishment will be satisfied.

But this means that the approach we are taking has drastically changed. While our theory began by focussing on whether or not an individual has agreed to a punishment practice, we have instead come to focus on whether or not an individual can expect to benefit under that practice. Should we be concerned that, within our purportedly contractarian theory, agreement has now apparently become superfluous? According to Simmons, the answer would be “yes.” He argues that under the consent tradition:

200

Obligations… not only require the performance of voluntary acts, but require that these acts be deliberate undertakings; an individual cannot become obligated unless he intentionally performs an obligation-generating act with a clear understanding of its significance.

As Simmons notes, the common thread between consent-based theories is the claim that “no man is obligated to support or comply with any political power unless he has personally consented to its authority over him.”201 This is because man is taken to have a natural right of freedom202—a freedom that can only be sacrificed by voluntary action.203 The attractiveness of theories like this is their ability to pay respect to the belief that “the course a man’s life takes should be determined, as much as possible, by his own decisions and actions.”204

200 Simmons, p. 64 (original emphasis).
201 Simmons, p. 57.
202 Simmons, p. 63.
203 Simmons, pp. 63-64.
204 Simmons, p. 69.
Thus, for Simmons, the ground of obligations must be the giving of a “‘clear sign’ of the appropriate sort to indicate the acceptance of the obligation and the transfer of the right.”\textsuperscript{205} The reasons for this standard are two-fold. First, claims Simmons, a deliberate undertaking is the only ground of obligation that ensures that “political obligations cannot be inherited or unwittingly acquired” and therefore “maximises protection of the individual’s freedom to choose where his political allegiance will lie.”\textsuperscript{206} There is an important sense in which we each have a right to determine our own lives—and this includes immunity from having unwanted benefits thrust upon us. Ensuring that an individual agrees to a punishment practice legitimates that practice in a way that merely pointing to an expected benefit cannot. Second, the requirement of a deliberate undertaking—a “clear sign”—lends “clarity and credibility” to the consent theory, as it is “as close to being an indisputable ground of moral requirement as anything is.”\textsuperscript{207} If, however, our theory focusses solely on whether an individual can expect to benefit from a practice, then we are eschewing this ‘clear sign’ requirement entirely.

I contend, however, that this is not as problematic as it would first appear. In order to show this, it is important to think back to the precise reasons why we saw an agreement-based theory as preferential to competing alternatives. I stated in Chapter 1 that this was largely because such a theory meets three important desiderata, namely that:

1. The theory makes an assessment of what is best for the agent; and
2. The theory makes this assessment in accordance with the agent’s subjective values; and

\textsuperscript{205} Simmons, p. 64 (original emphasis).
\textsuperscript{206} Simmons, p. 70.
\textsuperscript{207} Simmons, p. 70.
3. The theory makes this assessment on the basis of full information about the agent at the time of the assessment.

Now, as noted in Chapter 1, there may be some disagreement about whether or not these factors are important to assessing a theory of punishment. But this is irrelevant to the current problem. Instead, our question is as follows: If these three desiderata are the qualities that make an agreement-based theory attractive to the contractarian, does my theory’s current focus on expected benefit mean that it fails to meet these requirements?

I contend that it does not. Consider the first criterion: that the theory makes an assessment of what is best for the agent. This is still clearly the case. The very foundation of the contractarian theory that I have constructed is an assessment of which decision is likely to bring the agent an expected benefit. If an individual could not previously expect to benefit from a punishment practice, it will not—on the theory I have provided—be morally permissible to punish her under that practice.

Likewise, the second and third criteria—that the theory makes this assessment in accordance with the agent’s subjective values and on the basis of full information about the agent at the time of the assessment—are also met. Our hypothetical contracting agents are, by design, thick persons with subjective values. This has been deliberately incorporated into the very foundation of the second criterion of the contractarian test. In this way, we are not Rawlsian contractualists using hypothetical agreement to abstract away from real individuals. We are not telling people “yes, I know you don’t like this policy, but I’ll enforce it on you anyway since it is what you would agree to if placed behind a veil of ignorance and stripped of all of your
contingent characteristics.” Instead, we are simply trying to work out what people would agree to if asked. The lack of actual agreement is a mere formality.

My theory, then, still satisfies the three desiderata that gave us reason to prefer an agreement-based theory like contractarianism in the first place. I contend that, when seen in this light, much of the concern about the inherent lack of actual agreement should evaporate. When we talk of ‘expected benefit’, we are not making a paternalistic claim about what is best for an individual—then foisting that good on the individual regardless of whether they actually agree to it or not. Instead, we are examining their subjective values and calculating exactly what will bring the individual the greatest utility according to those values. The ‘greatest benefit’ will be the very thing that the individual would agree to if given the chance.

A theory based on hypothetical agreement (via expected benefit) still ticks the very same boxes that made an actual agreement-based theory so attractive in the first place. In addition, there is the added bonus that eliciting agreement from a large population of individuals will be much easier under a theory based on hypothetical agreement rather than actual agreement (be it express our implied through certain actions, or tacit through inaction). In light of this, it is unclear why we would not make this move.

Is the Threshold of Unanimity Too High?

Having demonstrated that our contractarian test still exhibits the characteristics that make an agreement-based theory of punishment attractive, we can now move on to consider a second concern for its practical implementation—namely, whether the previously mentioned threshold of unanimity is too high.
Up until this point, I have sought to establish an ‘ideal’ theory for the moral permissibility of punishment that provides us with an answer to two questions: (1) is a particular instance of punishment morally permissible? and (2) is a wider punishment practice morally permissible? I have argued that an instance of punishment will at least be permissible when the individual being punished previously (i) had good reason to agree to that punishment practice; and (ii) gave some kind of agreement to that punishment practice. In this chapter, I have argued that if hypothetical agreement is sufficient for agreement, then both (i) and (ii) will be met where it can be shown that an individual expected to benefit from the implementation of the punishment practice immediately prior to deciding to commit a crime (or being falsely accused of a crime).

In the previous chapter, I also claimed that a wider punishment practice will be morally permissible where the above test is satisfied for all individuals who are subject to the practice—that is, where hypothetical agreement is unanimous. The first reason for requiring this high bar was that punishment, as an institution, necessarily impacts on basic human rights. The second—and more fundamental—reason came down to the way in which the contractarian theory derives its normative claims about punishment. What grounds the moral permissibility of a punishment practice under the contractarian theory is the fact that the moral permissibility of any instance of punishment performed under that practice will stem from the fact that the very person being punished has previously given their agreement (in some form) to the practice. This can only be ensured where we are certain that every individual falling under a punishment practice agrees to the practice—that is, where agreement is unanimous.

Now, the question might be asked: why must every instance of punishment under a practice be justifiable on the basis of an agreement, in order for the wider practice itself to be
morally permissible? Surely we could tolerate some instances of punishment that are not justified in this way, so long as the vast majority of instances are? But such a query largely misses the foundation, and appeal, of the contractarian approach. A theory like utilitarianism will often make a trade-off like this, justifying a clear harm to a minority on the basis of a larger benefit to the majority. But this at least makes sense. The moral currency of utilitarianism is utility—and any action that maximises this currency will be deemed morally permissible. In this way, when an individual is harmed in the name of punishment, the utilitarian can articulate a justification that—in the context of her own theory—is compelling. To the punishee, they can say something like “yes, this punishment harms you, but—overall—utility is maximised by the existence of this practice.” If this is true, then sufficient grounds will have been given for the moral permissibility of the practice under the utilitarian approach.

In contrast, what grounds moral permissibility in the contractarian approach (and, I will argue in Chapter 7, makes it immune to the many objections raised against consequentialist theories) is the fact that each individual has a self-interested reason to submit to the punishment practice in question. Under this theory, the currency of moral permissibility is not simply utility, but agreement. While utilitarianism can justify harming one individual in order to bring a greater good to others, the same cannot occur in the context of contractarianism. It means nothing, for example, to say to an individual “yes, this punishment harms you, and you did not agree to it, but—overall—more people in society agree to it than not.” While this might be an accurate description of the society’s position with regard to the punishment practice, it does not articulate a justification that—in contractarian terms—will make it morally permissible to punish that individual. The only thing that can justify a particular instance of harm for the contractarian is
the fact that the very person being harmed has given her agreement. To claim that other individuals have agreed to the harm in question is morally irrelevant.

Some objectors will no doubt remain unconvinced. They will continue to argue that not every instance of harm requires the agreement of the individual being harmed. But, in light of the above, I hope to have demonstrated that this is not really an argument against the unanimity requirement of contractarianism, but rather an argument against contractarianism altogether. My claim is simply that, if we are convinced by the contractarian approach to punishment—that is, if we believe that harming an individual is only morally permissible where that individual gives some kind of agreement—then a morally permissible punishment practice will, by necessity, require unanimous agreement.

Suppose that I have successfully made this case, and that unanimity of agreement is required for a morally permissible punishment practice. If (per the preceding chapter) ‘agreement’ is satisfied by ‘hypothetical agreement’, and if (per this chapter) ‘hypothetical agreement’ is entailed by ‘expected benefit’, then our talk of ‘unanimous agreement’ instead becomes talk of ‘universal expected benefit.’ Thus, on my contractarian theory, a punishment practice will be morally permissible so long as it brings an expected benefit to everyone over whom it applies. This is the purist position. Unfortunately, it is not clear how many punishment practices could meet such a high threshold. I have all the way through this thesis based my discussions of utility and benefit on a subjective notion of what is best for a person. While this assisted us in avoiding a paternalistic theory, it now potentially gives rise to a wide range of scenarios in which universal expected benefit would be unattainable—leaving very few cases where a punishment practice is morally permissible under the theory.
Now, in many ways, this is not fatal to the theory. What I have proposed here is a counter to Boonin’s claim that punishment is never justified. Even if only a handful of punishment practices are justified under the high threshold of my theory, we will have provided a successful reply to Boonin. I want to go further than this, however. I want to show that the contractarian theory can be implemented in a way that is practicable, and therefore facilitates the implementation of a comprehensive range of meaningful punishment practices. In order to do this, we will need to move our focus from the question of “when is a punishment practice morally permissible?” to “what kind of punishment institutions should we implement?”

The most concerning problem cases for the contractarian theory are those individuals who are not expected to benefit under a particular punishment practice, even where the vast majority of the population will. Such individuals might include those who will go on to offend at an alarmingly high rate, those who stand to gain little from the promised benefit of deterrence, or those simply with value systems far outside the norm. It is these individuals who will deny a practice unanimity, and therefore prevent it from attaining the status of moral permissibility on the pure contractarian approach. I recommend that when dealing with these individuals, it is useful to split them into two separate classes based upon their epistemic status: cases of non-expected benefit that we know of, and cases of non-expected benefit that we do not know of. I will deal with the first of these categories in Chapter 6, suggesting a number of alternative strategies that might be used to bring these agents ‘into the fray’—that is, to demonstrate that they can expect to benefit from the punishment practice in question. In Chapter 7 I will instead turn to consider those non-expected benefitees of whom we do not—or cannot—know.
Having clarified these two issues, we can go on to demonstrate how the contractarian theory of punishment would operate in the real world. Specifically, we have distilled the contractarian test down to a single claim—namely, that:

It will be morally permissible to punish an individual under a particular punishment practice if, at some prior time, that individual could have expected to benefit under that practice.

Let us then consider how this theory will work in practice. Suppose that a small society is in the process of negotiating—or, potentially, renegotiating—their punishment institution. Specifically, they are considering what kinds of punishment they should adopt for four particular crimes: murder, grievous bodily harm, theft, and digital piracy. The society has, through some mechanism, decided to criminalise these activities, but now needs to establish what the state-sanctioned responses should be to actions that contravene these laws.

I have chosen to focus on these four specific crimes in order to bring out various nuances of the contractarian theory of punishment. Murder is (arguably) one of the most serious and harmful crimes, and thus most likely to justifiably attract the harshest penalty possible. If a particular kind of punishment cannot be justified for murder, then it is unlikely to be justified for any other lesser crime. Digital piracy, on the other hand, falls at the opposite end of the spectrum of harm-causation. Piracy does not involve physical harm to a person, nor does it—unlike other property theft—involves depriving victims of the property that was stolen.\(^{208}\) Between these two extremes lie both grievous bodily harm (a less harmful version of the violation of bodily

\(^{208}\) Though, of course, it may involve depriving the victim of income—a point to which I will soon turn.
autonomy that occurs with murder) and theft (a more harmful version of the violation of property that occurs with digital piracy).

Consider, then, what occurs when our model society turns to consider how it will punish the offence of murder. Thanks to a comprehensive investigation, legislators have a wide variety of information that will assist them in calculating the appropriate level of punishment for this crime. First, it is known that all individuals living within the society disvalue having their own life taken by another. Due to the subjective notion of value inherent to our contractarian theory, however, the strength of this disvaluation varies across members of the society. Some individuals—the “survivalists”—disvalue their own demise to such an extent that they will happily barter almost all other rights in order to protect themselves. Other individuals—the “pragmatists”—disvalue their lives being ended to some extent, but weigh this disvaluation against a number of other competing factors.

Second, legislators have a fairly comprehensive idea of how different classes of individuals can expect to fare under a number of different punishment practices. Some classes, for example, will have very little chance of coming to the attention of the punishment institution, and therefore will be largely unaffected by the severity of the chosen punishment, outside of the utility created by its deterrent effects. Other classes, such as offenders and—more importantly—recidivist offenders, will be hugely affected by any change in the kind and severity of the punishment.

Third, and finally, legislators—through a series of pilot studies, and comparisons with other jurisdictions—have reliable data on how the implementation of varying punishment practices will affect the murder rate within their society, as well as details concerning other costs associated with each of these different practices.
Armed with this information, legislators can then go about calculating the most suitable punishment for the crime of murder. Suppose that they start with the most severe sanction available: the death penalty. Legislators can begin with the expected utility that will be derived from the practice of executing murderers—namely, its ability to deter (and therefore reduce) instances of murder across the community. The legislators would, presumably, note that the practice leads to a significant reduction in the murder rate of a society compared with the implementation of no punishment at all. So far, so good. The legislators can then use this data and, combined with their knowledge of different citizens’ disvaluations of being murdered, begin to assess the relative utility that this deterrence holds for different portions of society. Some, like the survivalists, will receive a huge utility from this deterrence, while others, such as the pragmatists, will receive a slightly lesser utility. The remaining population will largely fall somewhere in-between.

Having done this, the legislators can then move on to consider the potential disutilities of implementing the death penalty for murder. First among these would be the collateral costs associated with this punishment practice. The death penalty is, for example, one of the most expensive punishments available—even more expensive than a lifetime of incarceration. Such costs are ultimately borne by taxpayers, and so must be weighed to some extent against the utility gained in the form of deterrence. There may also be other less obvious collateral costs, such as the deterrence of this crime resulting in the incentivisation of other crimes. In this case, for example, legislators might discover that while the murder rate is significantly reduced by the threat of the death penalty, the number of individuals killed by each murderer increases. This, presumably, would be a result of the fact that the penalty for murdering one victim is the very

---

209 In Texas, for example, a death penalty case costs an average of US$2.3 million—roughly three times the cost of imprisoning an individual in a maximum security prison for forty years. (Death Penalty Information Center, “Facts About the Death Penalty” <http://www.deathpenaltyinfo.org/documents/FactSheet.pdf>.)
same as for murdering several victims—a killer can only be sentenced to death once, so multiple murders cannot worsen her sentence. Those who have committed an act of murder become more likely to carry out additional murders—particularly in the hope of covering up their original crimes.

A second disutility to consider would be the cost of \( U_{Restrain} \) suffered by different individuals across the population. This will encompass any potential benefits that individuals may now be forced to give up in order to avoid being sentenced to the death penalty. Many people will have little to gain from committing acts of murder—but some will. The sole heir to an enormous fortune would, for example, stand to benefit greatly by killing off his aging benefactor. Likewise, an embezzling executive might avoid severe adverse repercussions by dealing with his whistle-blowing underling. Others may be able to forge lucrative careers as assassins or contract killers. But even of those who stand to benefit from committing murder, very few would actually have the real inclination to act so as to bring about this benefit. Nevertheless, some will. And of those individuals who would commit murder in a society without punishment, some of those will now restrain themselves from killing in order to avoid execution. Thus, these individuals suffer a disutility when compared with how they would fare under the status quo.

It may seem unusual to take these kinds of disutilities into account. Benefits gained from the commission of a crime—particularly a crime as serious and reprehensible as murder—certainly do not seem as though they should carry a great deal of moral weight when considering the permissibility of punishment. But this is an inevitable part of a theory that bases its moral assessment of practices on the basis of what an individual has good reason to agree to, and that—in turn—establishes this by examining the benefit that an individual can expect to receive.
While we might make all kinds of moral judgements about an offender receiving a benefit from such a serious crime, the fact remains that it still is a benefit. As such, this benefit will play a part in her decision-making process when it comes to what kinds of punishment practices she has reason to agree to. If the implementation of a punishment practice means that a potential offender will now have to restrain herself from some kind of beneficial behaviour (that is, behaviour that would have benefitted her), then this will be a disutility that is relevant to our contractarian calculation.

Put another way, in order to claim that an offender’s punishment is morally permissible, we must be able to show that she could expect to benefit from—and therefore had good reason to agree to—the wider punishment practice in question. As such, the disutility of restraint must (along with other relevant disutilities) be outweighed by the utility of deterrence. If it is not, then the offender would not have expected to benefit from the punishment practice, and therefore did not have good reason to agree to its implementation. If this is the case, then it will be morally impermissible to now punish that individual under that practice.

Many objectors will remain unconvinced. To claim that we should take the offender’s interests into account in the first place is a hard sell. But to formulate our test in such a way that it makes a special place for their lost ill-gotten gains may be a step too far. Nevertheless, I do not see this feature as a weakness of the contractarian theory. Rather, it is strength. Incorporating this factor into our theory means that when we do demonstrate the moral permissibility of a punishment practice, we will be making an exceptionally strong claim. We will be stating that it

---

210 The necessity of taking restraint into account stems from the fact that my current theory is a purely voluntarist one—the moral permissibility of a punishment practice resting solely on the agreement of affected parties. This is not to say, however, that a similar theory could not be constructed that intersects with a rights-based theory like those of Nozick or Simmons (see Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974) and A. John Simmons, *The Lockean Theory of Rights* (Princeton: Princeton University Press, 1992), thereby removing the requirement of taking restraint into account.
is morally permissible to punish an offender because he had good reason to agree to the wider punishment practice, and that he still had good reason to agree to the practice even in light of all of the benefits he might have obtained through the commission of his crimes. If ever there were a solid foundation for the moral permissibility of punishment, this—it seems—would be it.

Thus, with all of the data concerning utilities and disutilities in hand, the legislators could now begin to make broad assessments about how certain portions of society are likely to fare after implementation of the death penalty for murder. Legislators can begin by calculating how the practice is likely to affect those who do not expect to actually be punished under the practice—that is, those individuals who will receive what I referred to in Chapter 3 as Beneficial Punishment. As a reminder, these individuals will have good reason to agree to a punishment practice where:

$$[U_{\text{Deterrence}} + U_{\text{Restraint}} + U_{\text{Collateral}}] > U_{\text{StatusQuo}}$$

Legislators can now begin to plug values into this equation in order to assess the moral permissibility of the death penalty for murder. To be clear, this equation would not be completed once for the entire society. This is not a consequentialist calculus, where the net utility of $U_{\text{Deterrence}}$ is weighed against the net disutilities of $U_{\text{Restraint}}$ and $U_{\text{Collateral}}$. Instead, the contractarian theory—in its purest form, at least—requires that this equation must be satisfied for all individuals across society. And the utilities and disutilities of deterrence, restraint, and other collateral costs will differ from person-to-person. Now, as conceded earlier in this chapter, it would be too onerous to require such a detailed check of every single individual, so legislators might instead carry out this assessment for large strips of society simultaneously. They would
need to satisfy themselves that the above equation is met for the survivalists as well as the pragmatists, and for those who stand to lose out by restraining from murder as well as those who would never entertain the thought of committing such a heinous crime.

But even if all of these groups satisfy the above equation, this will only ensure that there is an expected benefit for those who do not expect to be punished. This, however, is only one subset of the society. There are also those who expect to be punished, and—as noted above—these individuals must also expect to benefit under the proposed punishment practice. We must therefore show that, for this portion of society:

$$[U_{\text{Deterrence}} + U_{\text{Sanction}} + U_{\text{RestRAINT}} + U_{\text{Collateral}}] > U_{\text{StatusQuo}}$$

In the case of the death penalty, it seems fair to say that the disutility $U_{\text{Sanction}}$ will be massive. If murder is considered the most serious crime, then having one’s life taken must be one of the most serious losses that an individual can suffer. In light of this, it is unlikely that the disutility of execution will be outweighed by the utility of deterrence for those individuals who can expect to be punished for murder at some point in the future (even without piling on the further expected disutilities of $U_{\text{Collateral}}$ and $U_{\text{RestRAINT}}$). There are certain far-out cases where such a balancing might occur. Say, in a society where the probability of someone being murdered as a child is close to 1, and where the death penalty is capable of deterring murder to the point where this probability becomes almost 0. In this case, the person who expects to be punished might actually be better off receiving the death penalty as an adult, since he will have at least made it to adulthood thanks to the punishment practice in question.
Suppose, however, that—prior to the implementation of any relevant punishment practice—the likelihood of being murdered in our imaginary society is much lower. Individuals have a relatively good chance of living long, happy lives, and a good chance of making it to adulthood. Imagine, further, that the implementation of the death penalty causes a far smaller reduction in the likelihood of being murdered. If this is so, then it seems clear that individuals who can expect to be punished for murder (either rightfully or wrongfully) cannot expect to benefit under a punishment practice of the death penalty for murder. These individuals will do worse under the punishment practice than they would under the status quo. As a result, they will not have good reason to agree to the punishment practice, and so will not hypothetically agree to its implementation. It will not, therefore, be morally permissible to punish these individuals under this punishment practice. This means that the practice will fail to receive unanimous hypothetical consent, and that it must—for this reason—be rejected.

But this is not the end of the matter. Having discarded the severest of penalties, legislators can now begin to consider less harsh alternatives. One such possibility might be life imprisonment. Suppose that legislators find that life imprisonment leads to a deterrent effect roughly equivalent to the death penalty. On average, however, each instance of life imprisonment costs less in the way of resources than a death sentence. These two factors alone will already see the left-hand-side of the Beneficial Punishment equation ($U_{\text{Deterrence}}$) increase, and the right-hand-side ($U_{\text{Collateral}} + U_{\text{Restraint}}$) decrease. Put simply, this means that anyone for whom the equation was satisfied when assessing the death penalty will also have the equation satisfied for life imprisonment.

But it is not these individuals on whom we need to focus. If an individual is expected to benefit under a punishment practice as harsh as the death penalty, then that individual is almost
certain to benefit under a less severe punishment practice as well (assuming that the lesser sanction also brings with it some proportionally enticing deterrent effect). Our attention must instead be on those individuals who did not expect to benefit under the harsher penalty. Can these individuals now expect to benefit under the lesser penalty? The answer will be mixed. The reduction in the disutility of being punished (assuming that life imprisonment causes less harm than execution) may be enough to tip the scales for some punishees who now have their harsh punishment outweighed by the benefits of deterrence. Chances are, however, that many offenders will remain in a position where they cannot expect to benefit under such a practice. If this is so, then life imprisonment must go the same way as the death penalty, and we must continue to work our way through alternative punishments in order of decreasing severity.

At some point we will, it is hoped, reach a kind and severity of punishment under which no individuals fail to expect a benefit. This might—to select a purely arbitrary punishment—occur when the legislators reach a prison sentence of twenty-five years without parole. At this level of severity the legislators would find that all of those who expected to benefit from harsher sentences would still receive a benefit from this penalty, but so too would those individuals who actually expect to receive the punishment at some point in the future.

Consider, then, grievous bodily harm (GBH)—that is, serious physical injury inflicted on a person by the deliberate action of another. Suppose that GBH is also universally disvalued throughout our model society. Like murder, GBH involves a serious violation of the bodily autonomy of an individual. This similarity provides us with a starting point. We know the utilities and disutilities in play will be very similar to that when considering the crime of murder. Ultimately, we must show for each individual that the utility of deterring GBH outweighs (1) the disutility of any collateral costs surrounding the punishment practice (such as the practical cost
of implementation, and the potential incentivisation of other crimes), (2) the cost of individuals forgoing certain benefits that could be gained from committing GBH, and (3) for those who expect to commit GBH, the cost of being sanctioned under the punishment practice.

Regarding (1), the practical costs will be very much the same as under murder. The death penalty, or life imprisonment, or a twenty-five year prison sentence cost the same regardless of the crime for which they are being administered. The potential to incentivise other crimes will be similar, also. If, for example, the state seeks to implement the death penalty for GBH, they may find a spike in the murder rate. If GBH attracts the most serious punishment available, then offenders may be tempted to go to any lengths to hide their crime—including the murder of an individual who, in other circumstances, they may have merely assaulted. Insofar as (2) is concerned, there will once again be myriad scenarios in which an individual might benefit from the employment of GBH. Intimidation is a powerful motivator, and many people stand to gain much from the use of bodily harm as a way of influencing others. These individuals may now have to forego these benefits in light of the implementation of a sufficiently severe punishment practice. Finally, concerning (3), the disutility of \( U_{\text{Sanction}} \) will—just like the disutility of \( U_{\text{Collateral}} \)—be largely the same as in the case of murder. A particular punishment will give rise to the same disutility for an offender, regardless of the crime for which she is being punished.

The disutility suffered by most victims of GBH is, however, significantly less than that suffered by most victims of murder.\(^{211}\) This means that individuals will usually stand to gain less from the deterrence of this crime over more serious crimes like murder. As such, the left-hand-side value of the contractarian calculation (\( U_{\text{Deterrence}} \)) will be lower, meaning that the right-hand-side value (\( U_{\text{Collateral}} + U_{\text{Restraint}} + U_{\text{Sanction}} \)) must also be reduced. This will necessarily require that

\(^{211}\) I say “most” as there are conceivable scenarios in which a person in certain circumstances, and with particular values, may stand to suffer more disutility by being the victim of GBH than the victim of murder.
we adopt a kind and severity of punishment that is somewhat less than that prescribed for murder. We might find, for example, that we do not hit the threshold of universal expected benefit until we arrive at a prison sentence of ten years without parole.

What allows us to reach satisfactory conclusions on the issues of murder and GBH is the fact that—by stipulation—all citizens disvalued the prospect of suffering a violation of their bodily autonomy in these ways. It is these disvaluations that give deterrence a utility sufficient enough to outweigh the disutility of being punished. But what happens in the case of crimes where there are no such widespread—and strongly held—disvaluations?

Property crimes perhaps give us our best examples of these. There are those, for example, who see no value in personal ownership and go so far as to claim that “all property is theft.” Finding unanimous agreement regarding these crimes (that is, property crimes which actually deprive the original owner of the stolen property) is hard enough. But consider how much more difficult it will be to garner agreement for property crimes that do not involve such deprivation. Take, as an example, digital piracy—that is, the illegal copying and distribution of digital movies, music, games, and other content. There will clearly be some individuals within our society who strongly disvalue piracy. Chief among these will be those who create and produce such content, and rely on copyright laws to ensure that they receive fair remuneration for their work. But there will also be a portion of society that disvalues digital piracy to a much lesser extent. They see why content creators have reason to object to this kind of behaviour, but find this largely unmoving insofar as their own situation is concerned. When considering punishing digital piracy, these individuals are unable to find any satisfactory answer to the question “what’s

212 To be clear, there may be some individuals out in the world who are either (1) indifferent to the prospect of losing their life, or (2) actively value losing their lives. These individuals are strange ones, but not entirely inconceivable. For this reason, I will turn to consider how we should deal with such people in the following chapter.
in it for me?” How, then, can any sort of punishment practice be permissibly implemented over such a population?

First of all, we can note that while the “expected benefit” requirement is based on subjective valuations, it also relies upon objective facts—some of which may be unknown to an individual. In the context of digital piracy, for example, there may be ways in which piracy does harm an individual without their knowledge. We might, for example, point to the fact that the price of legal content will rise in order to compensate for the effects of piracy. In this way, non-content producers still have good reason to disvalue piracy (assuming they in turn disvalue having to pay more for a particular product). Unfortunately, this argument will not work on the pirates, as they have no intention of paying for legal content in the first place. But in order to address these individuals, we can point to an alternative argument: We could note that piracy results in a stifling of the creative industry, and the production of less content overall. Assuming that the pirate values a wide variety of content, then this will give him at least some reason to disvalue piracy.

Having informed people’s valuation in this way, legislators could then go on to carry out a similar assessment of punishment practices as they did with the case of murder. They would start with the most severe penalty, then begin to reduce this until they reached a severity of punishment that would ensure that all individuals could expect to benefit. We need to remind ourselves that—as previously stated—a punishment practice will only be permissible where it brings an expected benefit to all to whom it applies. Severe punishment for these crimes will not benefit those who place a minimal value on its deterrence, as the potential harm of being punished for the offence will outstrip the foreseeable benefit of deterring potential offenders. In this way, more serious forms of punishment—while endorsed by those who place a high value on
the deterrence of the offence—will fail to provide a universal expected benefit. Ultimately, this will see legislators focussing primarily on those who both (1) disvalue digital piracy to the least extent, and (2) can expect to be punished under the practice. It is these individuals to whom it will be most difficult to ensure an expected benefit.

As we follow this approach, it soon becomes clear that a large portion of the calculations we have carried out are unnecessary. As I have noted: once it is shown that an individual can expect to benefit under a particular punishment practice, it can be presumed that that same individual can expect to benefit under all less severe punishment practices (provide that sufficiently attractive deterrence is still achieved). Ultimately, it is a matter of reducing the severity of the punishment practice until all individuals can expect to benefit under the practice.

What we need to do, then, is focus on the set of individuals who can expect to benefit least from punishing a particular crime, then reduce the severity of the punishment until even this group can expect to benefit under the practice. Thus, in our practical application of the contractarian theory of punishment, we have arrived at something like Rawls’s maximin principle.\textsuperscript{213} Stated in its simplest terms, the contractarian theory holds that a punishment practice will be morally permissible only where it maximises the expected benefit of those who are expected to benefit least. If these individuals can expect to benefit more in a world without the punishment practice, then that practice must be discarded in favour of another alternative (including the implementation of no punishment practice at all).

At this juncture, it is worthwhile taking stock of our current position. At the outset, I claimed that the initial appeal of the contractarian approach was that it allowed us to give a moral basis for punishment that can be justified even to the individual currently undergoing punishment (i.e., that he had reason to agree to the punishment practice because he could have expected to benefit under the practice). Ultimately, however, this approach has led us to the position outlined in the previous paragraph—one in which a punishment practice will be morally permissible only where it maximises the expected benefit of those who are expected to benefit least.

But there are two foreseeable—and potentially serious—problems with this approach. First of all is a concern that arises surrounding the ‘ratcheting down’ of the severity of the punishment as prescribed by the contractarian theory. In order to make a punishment practice attractive (and therefore agreeable) to offenders, we must gradually reduce the severity of the proposed punishment. But this, in turn, also reduces the amount of deterrence provided by the punishment, and therefore makes offenders more likely to commit the crime in question. Even more problematically, there is a certain severity of punishment below which any deterrence is lost. These ‘slap-on-the-hand’ type punishments provide no utility, and therefore give individuals (especially expected offenders) even less reason to agree to the punishment practice in question. So, insofar as the contractarian is concerned, all punishments below a certain level of harm are essentially the same as having no punishment at all. If our theory cannot justify any punishments above this level, then our theory results in the same practical outcome as Boonin’s—that is, that no (effective) punishment practices are morally permissible.
Ultimately, however, this simply means that we have a reason internal to our theory as to why we need to search particularly hard for some justifiably harsh penalties. This will be in contrast to most scenarios in which calls for harsher penalties depend on some external reason like “I just think that this crime deserve a harsher punishment.”

The second problem—and one that may interfere with our ability to justify these harsher penalties—is the apparent veto that has been afforded to those who are worst-off. On the view I have outlined so far, if even one individual cannot expect to benefit from a punishment practice, then it will be impermissible to implement that practice over the wider population, regardless of the expected benefit that might be afforded to every other citizen. This, of course, is what distinguished the contractarian theory from a more familiar consequentialist approach. But nevertheless, it is a difficult bullet to bite. Especially given the fact that the worst-off individuals who are given this veto are those individuals who are likely to offend the most often, and to the greatest degree. In the next chapter, I will delve into this problem further. While any further relinquishment of the unanimity requirement would weaken the contractarian theory to an unacceptable degree, I will suggest several strategies for bringing potential-veto holders “into the fray.” That is, showing that on a certain interpretation of their interests, they can expect to benefit from the punishment practice in question.
Chapter 6: Problem Cases for a Contractarian Theory of Punishment

So far, I have argued that my contractarian theory of punishment boils down to one primary principle: that it will be morally permissible to punish an individual under a particular punishment practice if, at some prior time, that individual could have expected to benefit under that wider practice. I have also argued that, on the ideal version of the theory, a punishment practice will only be morally permissible when it meets this requirement for all individuals to whom it applies. Put another way, a morally permissible punishment practice will be one that confers universal expected benefit on a population.

The inevitable difficulty with this universality requirement is that a punishment practice can apparently be rendered morally impermissible by the existence of a single individual who could not have expected to benefit from that practice. The ‘worst-off’—that is, those who could not expect to benefit under the practice—are given a kind of veto over the practice being implemented over the wider population. Unfortunately, this might see the disqualification of many otherwise useful punishment practices. Ideally, we want to maximise the punishment practices that we can implement—so it is worthwhile considering whether such vetoes would occur, and to what extent they would be problematic. Ideally, we want to demonstrate that many of these non-expected-benefiters can be brought “into the fray”—that, upon further investigation, these do in fact benefit from the punishment practice in question. In this chapter I will examine three potential kinds of alleged non-expected-benefiter—the Value Outlier, the Vigilante, and the Foole—and discuss potential strategies for responding to each.
The First Case: The Value Outlier

The theory outlined in the preceding chapters assumes that the contractarian approach can, in principle, provide justification for a single, community-wide institution of punishment—that is, an institution comprised of many universal punishment practices directed towards a variety of different crimes. We must consider, however, whether such an assumption is reasonable in light of the fact that different individuals both (1) have different expectations about the likelihood of different crimes being committed, and (2) attach different amounts of value to being protected from these crimes.

It is easiest to deal with the former of these concerns—that is, that different people have different expectations about the likelihood of certain crimes being committed—first. This is because these differences in expectations can largely be reduced down to a paucity of information—something which our contractarian theory is especially designed to account for. Whether or not an individual has reason to agree to a particular punishment practice will depend (among other things) on the deterrent effect of the punishment. Effective deterrence will result in a quantifiable reduction in the likelihood of an individual being harmed by an offence, and this reduction will confer a utility that will provide that individual with a reason to agree to the practice in question. It is true that individuals may have (potentially diverse) expectations regarding this deterrent effect, but the fact remains that the actual reduction in the likelihood of an individual being harmed will be an objective and concrete value. Our contractarian theory bridges this divide between expectations and actuality by ensuring that hypothetical contractors are given “full knowledge of the relevant utilities and disutilities of the [punishment] practice” before making their decisions. This will include highly accurate information about the rate at
which offences will occur both pre- and post-implementation. While an individual might feel more or less safe by agreeing to a certain punishment practice, this will have no bearing on whether or not they have good reason to agree to that practice. Instead, the contractarian test requires real expected benefit. In this way, differences in expectations about how likely people are to commit crimes has no part to play in calculating whether or not an individual can actually expect to benefit after the implementation of certain punishment practices.

This, then, deals effectively with those individuals who suffer from a lack of knowledge regarding the effectiveness (or non-effectiveness) of certain punishment practices. What is far more problematic, however, is the observation that different individuals attach different amounts of value to being protected from different crimes.

One foundational premise of my contractarian theory of punishment is the claim that for every person, there are certain things that are valuable for that person. I stated in Chapter 2 that it did not matter what the specific source of these values was. However, for the purposes of my argument I have adopted a subjective theory of value. There were several reasons for this—chief among which is that a subjective theory softens some of the paternalistic overtones of looking to hypothetical (not actual) agreement. But there is one severe downfall of a subjective theory of value—namely, that individuals may come to value very different things. The benefit of deterrence is not only dependent on the concrete reduction in the likelihood of a person being harmed by an offence, but also on the value that that individual attaches to this reduction. For many offences, there will be wide agreement. Most of us, for example, attach an extremely high value to not being the victims of murder. For other offences, however, agreement may not be so easily found. Such divergence greatly increases the chance of there being individuals who fail to expect a benefit under a proposed punishment practice.
One response to this problem is to posit some objective theory of value—that is, to say that while certain individuals may *not* value the non-commission of certain offences, they at least *should*. This, however, would require the formulation of a convincing argument for value objectivism. While the perfectionist reader may be easily convinced, those who subscribe to theories such as hedonism or desire satisfactionism will be far more difficult to persuade. I therefore plan to double-down on the subjective theory of value that I have employed so far. In order to do this, however, we must address the issue of those individuals whose values diverge greatly from the common values of their shared community. These “value-outliers” fall on a spectrum, ranging from those who (1) are merely *indifferent* to what the majority disvalues; to (2) those who actively *value* what the majority disvalues.

While a simple, singular solution to all cases on this spectrum would be preferable, it is unlikely that such a tidy response can be given. Instead, I will consider a range of strategies for addressing each of these kinds of value-outlier in turn. To be clear, these strategies will not succeed in bringing all value-outliers into the fray. Some will remain outside the reach of our theory, and the existence of such individuals *will* (due to the unanimity requirement) place a limit on the selection of punishment practices that it is morally permissible to implement. But—as stated at the beginning of this chapter—the purpose of these strategies is not to provide a sure-fire way to garner *every* individual’s hypothetical agreement. Rather, they are presented as a way of ensuring that we have not been too quick to judge which individuals will, and will not, give rise to problematic vetoes over otherwise useful punishment practices. Our intention here is to gain a picture of the widest possible scope of individuals who can be taken as expecting to benefit under each punishment practice, before making any judgements about whether or not the theory provides us with a sufficiently comprehensive institute of punishment.
Consider, first, the individual $P$ who is merely indifferent to the commission of a certain offence $O$ that is widely disvalued by other members of his society. Put another way, $P$’s values are such that her life will fare no better, nor worse, should $O$ be committed (even if $P$ is the victim of $O$). Perhaps $P$ lives in a puritanical society where exposing one’s head in public is seen as a serious wrong. $P$, however, could not care less about whether a person’s head is covered or not. In light of this, $P$ will have no interest in deterring the behaviour in question, and therefore cannot expect to benefit from the implementation of a punishment practice $X$ that deters the commission of $O$. This, in turn, will mean that at no point in time will $P$ give her hypothetical agreement to $X$. As such, it will be morally impermissible to implement $X$ over $P$, and—due to the unanimity requirement—also impermissible to implement $X$ over $P$’s wider society.

This is problematic. While the example of a punishment against bare heads is trite, we can imagine circumstances in which a more reasonable punishment against a more serious crime is prevented from implementation due to the existence of some indifferent value outlier. Perhaps, for example, we encounter one of those individuals from the previous chapter who rejects the concept of personal property and, as such, is utterly indifferent to the commission of theft. As the contractarian theory stands, it seems that the existence of this individual would mean that her society was entirely unable to implement a morally permissible punishment for theft. This would be a difficult bullet to bite, so it is worth considering whether it is a necessary entailment of our theory.

I would argue that the contractarian approach has sufficient resources to circumvent problem cases like this. Specifically, it provides us with the opportunity to offer $P$ a reason for agreeing to implement $X$ that is external to the standard utility versus disutility calculus that we have focussed on so far. The argument might go something like this: Just as $P$ is indifferent to an
offence that others disvalue, so too will some others be indifferent to offences that $P$ disvalues. Given this, $P$ may very well want to strike a bargain with his contracting partners. Specifically, $P$ will have good reason to agree to punishments for offences to which he is indifferent, but that others disvalue, so long as others will agree to punishments for offences to which they are indifferent, but that are disvalued by $P$. In this way, each party will have a better chance of securing punishment practices that deter those offences that he or she disvalues.

Far more troubling is the case of the individual who actively values the commission of certain crimes. What do we say, for example, to someone who—for masochistic reasons or other—values being seriously assaulted or murdered? One move we might make is to point to certain inconsistencies in this agent’s set of values. Suppose, for example, that he also values seeing his infant child graduate from college. Such an end seems inconsistent with a complete disregard for the preservation of his own life. Given this, we could argue that the individual in question does in fact value not being killed, and has reason to agree to a punishment practice that deters others from taking his life.

But what about those individuals who manage to ensure consistency across their values? We might be able to deal with some of these cases by ensuring that our rules of punishment are worded in suitably nuanced terms. Thus—to take the assault example—instead of attempting to implement a punishment for serious assault, we could instead attempt to implement a punishment for “interfering with another person’s body in contravention of their desires.” Such a punishment practice will still achieve the deterrence goals of those who disvalue being seriously assaulted. The advantage of this reformulation, however, is that it may now elicit the agreement of some masochists, as it no longer expressly punishes the behaviour that they value. But a number of masochists may still fail to have reason to agree to even this reframed punishment practice.
Consider, for example, the individual who actively and consistently values non-consensual interference with each other’s bodies (including his own). Such a person is a strange one indeed—placing no value whatsoever on his continued safety, nor survival. Nevertheless, he is a conceivable agent, and his lack of expected benefit from any punishment for bodily interference means that—on the contractarian approach—it will be morally impermissible to implement punishment practices for even the most basic and widely accepted punishment practices.\(^{214}\)

I contend, however, that we might employ the very same solution that we used to address *indifferent* value outliers. While the extreme masochist’s values may mean that he has no reason to agree to a specific punishment for bodily interference, his values may be such that he has reason to agree to other kinds of punishment—and wish to elicit the agreement of others for *these punishments*. Thus, he might decide to accept the burden of agreeing to a punishment practice that will bring him an expected *disutility* in order to elicit the unanimous agreement of others to punishment practices that will bring him an outweighing *utility*.

Essentially, our error thus far has been in considering each punishment practice on a solitary basis. A punishment system comprises a set of punishment practices, and when considering which of these practices an individual has reason to agree to, reasons other than the expected benefit of each practice taken alone may come into play. The total utility of the punishment system is, for each individual, greater than the sum of the utilities of its parts. Thus, at the hypothetical level, our contractors need to not only consider the raw expected utility or disutility of each specific punishment practice. They must also consider (1) the utility derived from other punishment practices within the system; and (2) the ramifications that rejecting *this* punishment practice (which may bring with it an expected disutility) will have on the likelihood

---

\(^{214}\) Here I have in mind any kind of punishment for assault, sexual assault, and murder.
of the implementation of other punishment practices (which may bring with them outweighing expected utilities).

To be clear, we are not moving our discussion to focus on the justification of the punishment institution as a whole. Our focus is still on justifying each punishment practice on a case-by-case basis. What we are acknowledging, however, is that a benefit of agreeing to an otherwise non-beneficial punishment practice may—on occasion—be the elicitation of the agreement of other contractors to other beneficial punishment practices. I will henceforth refer to this as the **Collateral Contractual Benefit**.

All of this is to simply say that we now have an additional utility that should be included in our contractarian calculus. Initially, we claimed that the utility values for Beneficial Punishment and Detrimental Punishment could be given by the following equations:

\[
U_{\text{BeneficialPunishment}} = U_{\text{Deterrence}} + U_{\text{Restraint}} + U_{\text{Collateral}}
\]

\[
U_{\text{DetrimentalPunishment}} = U_{\text{Deterrence}} + U_{\text{Sanction}} + U_{\text{Collateral}}
\]

Now, however, we can revise these equations to appear as follows:

\[
U_{\text{BeneficialPunishment}} = U_{\text{Deterrence}} + U_{\text{CollateralContractualBenefit}} + U_{\text{Restraint}} + U_{\text{Collateral}}
\]

\[
U_{\text{DetrimentalPunishment}} = U_{\text{Deterrence}} + U_{\text{CollateralContractualBenefit}} + U_{\text{Sanction}} + U_{\text{Collateral}}
\]
It is perhaps easiest to see the importance of a Collateral Contractual Benefit by way of an example: Suppose that there was a society made up of equal parts cat and dog owners. That is, half of the population own one or more cats, but no dogs; while the other half of the population own one or more dogs, but no cats. Imagine, further, that this society had a comprehensive range of laws protecting these cats and dogs from harm—even harm caused accidentally by others.

The time now comes to establish what kinds of punishments should be attached to these crimes. Cat owners will, presumably, see themselves as expecting no benefit from any kind of punishment for the harming of dogs. In fact, it is highly likely that most cat owners will stand to suffer a mild detriment from any such punishment practice. They can expect to suffer various disutilities—such as the collateral costs of enforcing and administering any punishment for dog-related crimes, as well as the cost of sanction if—for example—a cat owner happens to accidentally run-down someone’s dog. The inverse will be true of dog owners when it comes to punishment practices for cat-related crimes. Since, on the contractarian theory, a punishment practice will only be morally permissible where everyone can expect to benefit, this apparently leads to something of a stalemate.

But suppose that our hypothetical contractors are a little more sophisticated than this. Imagine that cat owners calculate that the detriment they could expect to suffer from certain dog-related punishment practices would be hugely outweighed by the benefit they could expect to receive from the analogous cat-related punishment practices. They also realise they would be happy to agree to a number of dog-related punishment practices if they knew that this would entail the implementation of a similar set of cat-related punishment practices. Further, they deduce that dog owners will have come to the very same conclusion.
In light of this, both cat owners and dog owners will each have reason to bear the burden of punishment practices from which they do not expect to benefit in order to elicit the implementation of punishment practices from which they do expect to benefit—that is, to gain certain Collateral Contractual Benefits. It is this reciprocal trading of implementations that can provide reason for a self-interested rational individual to accept certain otherwise non-beneficial punishment practices.

This approach will not necessarily lead to a determinative bargain, but potentially a range of bargains—all of which are capable of providing a moral justification for the punishment practice in question. Consider an analogous case: Suppose that I am accepting offers for the purchase of my car. Ideally, I want to receive $10,000—but I would be happy to accept as little as $8,000. Jack is interested in purchasing my vehicle. Ideally, he will spend around $7,000—but he is willing to stretch his budget as high as $9,000. What, then, is the optimal bargain for Jack and I to agree upon? It seems, in fact, that there are a range of acceptable outcomes—namely, any sale price between $8,000 (the lowest price I am willing to accept) and $9,000 (the highest price Jack is willing to pay).

The same will occur in cases of hypothetical agreement regarding punishment. To return to the cat and dog example from above, there may very well be a range of cat-protecting punishment practices (differing in kind and/or severity) that all dog owners are willing to accept in order to see the implementation of a particular dog-protecting punishment practice $PPD$. Of these, there may be some smaller sub-set that all cat-owners, in turn, consider a worthwhile price for the implementation of $PPD$. This smaller sub-set represents a set of bargains that the entire population of both cat and dog owners will have good reason to agree to. For this reason, the
implementation of any one of these punishment practices will be morally permissible (provided that PPD is simultaneously implemented as well).

But while this approach may draw many more contractors in to the fray, some will continue to remain excluded. There may be individuals who strongly believe in punishing only what is minimally necessary or who value a very limited number of things. As such, they have little to gain from either the wider institution of punishment, or the diverse range of punishment practices that comprise it.

What can the contractarian theory say in reply to such individuals? Unfortunately, the answer seems to be “not much.” I do not, however, believe that this counts as a fatal blow for the theory. It is worth noting that we are not here talking about mere psychopaths. Many cases of psychopathy are accompanied by a desire for self-preservation that would give the psychopath good reason to accept the implementation of certain punishment practices. The individuals that provide a problem case for our current theory of punishment fall even further outside of societal norms. They are so extreme—so outside the regularities of normal human values—that no amount of labyrinthine reasoning can bring their values to coincide with those of their fellow citizens. As such, it seems that there may be good justification for simply excluding these individuals as rational contractors. This exclusion comes at a price: specifically, all contractarian justification for punishing these individuals evaporates. But punishment is not the only tool with which we can respond to these extreme outliers. When such an individual offends, we may still be justified in utilising other non-punishment methods such as rehabilitation or isolation.

---

215 Even if the psychopath has no intention of wilfully surrendering to these punishments himself—more on this below.

216 It might be argued that isolation can cause harm to the offender, and so should count as a case of punishment. We must remind ourselves, however, that—under Boonin’s definition of punishment—harm is necessary, but not sufficient for something to count as an instance of punishment. The further vital requirement is that this harm is intentional, not accidental. No doubt there are some cases of isolation where harm is the primary intention (consider, for example, the use of solitary confinement in prisons), and these cases can rightfully be deemed examples of
The Second Case: The Vigilante

The Vigilante presents us with a special case of an individual whose values differ from those of her wider society. Unlike the Value-Outlier, the Vigilante may very well disvalue the same kinds of behaviours to the same extent as the rest of her community. Where she differs, however, is in her values regarding the administration of punishment for wrongdoing. She does not believe that such punishment should be administered by a central authority, but instead seeks a system of personal retribution—of vigilantism. Suppose that she is a widow and that she hunts down and executes her husband’s killer. She does not do this with the primary intention of bringing about deterrence, but instead acts from a deep-seated notion of retribution. She assigns great value to personally harming the killer, regardless of whether it has any wider deterrent effect. In fact, in many cases, a vigilante may seek personal retribution even where it comes at the cost of some level of deterrence. What can the contractarian say in response to such individuals?

If the vigilante truly values personal retribution, then practical rationality will—all things considered—hold that she has good reason to pursue this course of action. But this does not mean that contractarianism will endorse a wider practice based upon this behaviour. It is important to remember that under the contractarian theory, a punishment practice will be morally permissible only where it brings about a universal expected benefit (thereby entailing unanimous hypothetical agreement) to all who might be subjected to its sanctions. This expected benefit must even extend to those who expect to be punished, and will come in the form of the increased punishment. What I have in mind here, however, are cases of isolation where the primary intention is instead the protection of innocent individuals. We can identify these cases of (non-punitive) isolation by considering the scenario in which an offender’s isolation benefits him on the whole, and asking whether the isolators would see this as a failure. If they would not, then this will count as a case of non-punitive isolation, and would stand as a justifiable alternative to punishment.
protection afforded by the deterrent effect of the punishment practice in question. Put simply, the detriment an offender might expect to suffer in the form of punishment must be outweighed by the benefit that the wider practice provides her in the form of protection. If vigilantism does not have a deterrent effect, then there will be nothing to provide an expected offender with a reason to agree to such a practice. The individual who expects to be subjected to vigilante justice receives no benefit and will therefore be worse off than he would be in a world without such a system. Given this, unanimous hypothetical agreement will not be obtained and the punishment practice will fail to be morally permissible under the contractarian approach.

But perhaps our vigilante is more community-minded. Perhaps she engages in personal retribution with the very purpose of bringing about deterrence, knowing that her actions will discourage potential offenders. The contractarian theory of punishment endorses punishment practices where they provide an expected benefit to all in the form of deterrence—so what if a system of vigilantism was also capable of bringing about deterrence? This is a bullet that the contractarian must bite. An individual will have reason to agree to a system of vigilantism if it brings about a level of deterrence such that the individual can expect to benefit under this system. If this expected benefit is greater than that brought by any other system (including any system involving punishment meted out by a central authority), he will have reason to prefer vigilantism over all competing alternatives.

But how likely is it that a system of vigilantism will be preferable in this way? I contend that personal retribution is simply not as effective at deterrence as a punishment that is administered by a central authority. There are a number of features that a contractarian-endorsed

---

217 It might be argued that he gains the bonus of being able to himself pursue personal retribution against others—but this benefit is contingent on him being the victim of a crime in the first place. He may go his entire life without having an offence committed against his person, and therefore is never able to take advantage of this benefit. The benefit afforded by deterrence-based punishment is, instead, guaranteed and automatic.
system of deterrence requires. First, we want punishments to be proportional and restrained. If they are not—if, for example, the individual handing out justice has no limits on the harm he can cause—we will rarely be able to establish the expected benefit that will give us reason to agree to such a system. We will instead take our chances in a world without deterrence, for fear of receiving a grossly disproportionate amount of harm in response to a trivial offence. Second, we want punishments to be directed towards the actual offenders. In order for a punishment to act as any sort of deterrence for an offence, the punishment practice must operate so that committing the offence in question will actually increase the offender’s likelihood of being punished. If a punishment practice is arbitrary—if it is just as likely to punish the innocent as the guilty—then an individual’s decision to offend will have no bearing on the probability that he is punished, and thus the punishment will no longer have any kind of deterrent effect. Third—and relatedly—a punishment system will be most effective where it is built upon a system of comprehensive and thorough investigation, making it more likely that offenders are accurately identified and punished.

These then, are some of the virtues of an effective punishment system: proportionality, restraint, accuracy, and an ability to engage in comprehensive and thorough investigation. Where vigilantism can better display these virtues than a centralised institution of punishment, we will have reason to prefer the former. I contend, however, that this rarely—if ever—will be the case. A well-run justice system in cooperation with a transparent policing service will almost always be better equipped to carry out the task of punishment, and therefore operate the most effective form of deterrence. While the vigilante may have a deep desire to impose his own

\[^{218}\text{There are many other potential virtues that we would expect an ideal punishment system to possess—but I mention these as only a representative sample, and as among those that have the greatest bearing on the system’s ability to create deterrence.}\]

\[^{219}\text{At least in principle. I remain silent as to what extent our current justice system meets this standard.}\]
form of justice, he will benefit more from allowing the appropriate authorities to do this on his behalf.

**The Third Case: The Foole**

Suppose we know that, at the appropriate time of assessment, an individual $P$ can expect to benefit from a certain punishment practice $X$. If this is true, then according to the argument I have presented so far, $P$ will both (i) have good reason to agree to $X$; and—for this reason—(ii) hypothetically agree to $X$.

But this is only a claim about an individual having reason to initially *enter into* the agreement. So far, we have failed to explain why an individual will have reason to *continue to adhere* to that same agreement over time. While it may be true that we have good reason to agree to certain punishment practices, what is to stop an individual agreeing to these practices while he expects to benefit from them (that is, while he reaps the utility of their deterrent effects), then defecting when they no longer do? Specifically, what reason does an individual have to continue to agree to a punishment practice when he becomes the very subject of its sanctions? If $P$ has good reason to agree to $X$, then this is because the expected utility (in the form of deterrence) outweighs the expected disutilities (such as the potential harm of being punished). But if $P$ already has reason to agree and adhere to $X$, it seems he will have even greater reason to become a party to $X$—thereby obtaining its promised utility—but defect when it is *his* turn to suffer punishment—thereby eliminating the risk of any suffered disutility.
Michael Bratman provides a neat analogy of this concern when considering the rationality of fulfilling assurances:

I am faced with a choice whether sincerely to assure you at \( t_1 \) that if you help me at \( t_2 \), I will help you at \( t_3 \). If I do sincerely assure you at \( t_1 \) then, if you assume I would follow through if you were to help me, you can judge that your life goes best if you do indeed help me at \( t_2 \). But if I do not sincerely assure you at \( t_1 \) you will reasonably believe that you do better not to help me at \( t_2 \), since you then will reasonably not expect that I would reciprocate at \( t_3 \). My life goes better if we each help the other than if neither helps the other, and I can ensure that better outcome by sincerely assuring you. But it is also true (we are to suppose) that at \( t_3 \) if you have already helped me at \( t_2 \), my life goes even better if I then abandon my assurance and do not help you. So it is not clear that even if I settle on this assurance plan at \( t_1 \), and you do indeed help me at \( t_2 \), it will at \( t_3 \) be rational for me to follow through with my plan.

Usually when I make an agreement, I do so while expecting some benefit (deterrence, in the context of punishment) and understanding that I will—or may—have to pay some cost in the future (receiving punishment for any crime I might commit). But after receiving the expected benefit of deterrence, what reason do I have to fulfil my obligation of allowing myself to be punished? What reason do I have not to defect from the agreement?

In order to make sense of this objection, it is necessary to establish exactly what “defection” amounts to in the context of the contractarian theory of punishment. To be clear, to defect from the punishment agreement is not to commit the offence in question. An agreement to implement a particular punishment practice is not an agreement to refrain from offending. Instead, it is an agreement to punish (and receive punishment for) a particular offence in a

---

particular kind of way. To defect is therefore to *avoid* the punishment prescribed by a practice that one has previously agreed to.

But even this is too general—for there are two very different ways in which one can ‘avoid’ punishment. The first of these ways is to simply *resist* punishment. This would cover behaviour such as hiding the evidence of your crime, deliberately eluding apprehension by law enforcement, escaping from prison, or repeatedly refusing to fulfil court mandated punishments like the payment of fines or the completion of community service. The second way of avoiding punishment involves an offender actually *revoking* the punishment agreement that one has previously agreed to. This is an incredibly important distinction, as the first of these methods of avoidance does nothing to undermine the punishment agreement itself. Even if an offender makes it difficult for officials to punish him, this does nothing to nullify the fact that there is an enforceable agreement still in effect between the individual and the state that provides a moral justification for the state harming the individual. The offender is permitted resistance without this affecting the legitimacy of the punishment agreement. Even Hobbes acknowledges that, despite any prior agreement, “a man cannot lay down the right of resisting them, that assault him by force. To take away his life,” and that the same can be said of “Wounds, and Chaynhs, and Imprisonment.”

However, avoiding punishment in the second way—that is, by revoking the punishment agreement altogether—is far more serious. This is because such an action *does* undermine the very legitimacy of the punishment practice. How might this revocation actually occur? In Chapter 2, I argued that an individual *P* will have reason to agree to a punishment practice where she can expect to benefit under that practice—that is, where she can expect do better under the punishment practice than she would under the status quo. Such a calculation is made by

---

221 Hobbes, Part I, Chapter 14, p. 192.
comparing the gained utility of the practice (in the form of $U_{\text{Deterrence}}$) against the potential disutilities of the practice (in the forms of $U_{\text{Sanction}}$, $U_{\text{Restraint}}$, and $U_{\text{Collateral}}$). Now, however, we must consider the scenario in which an individual is able to obtain that utility while avoiding some—or perhaps all—of the disutilities. In practical terms, this will be the scenario in which (1) $P$ agrees to the punishment practice $X$, (2) $P$ commits the offence covered by $X$, and (3) $P$ subsequently withdraws her agreement to $X$ after having committed the offence. This is what true defection amounts to in the context of the contractarian theory of punishment. It will also—in certain suitable circumstances—be the scenario that brings the individual the greatest possible actual benefit. What, then, is to stop the individual having the very same reason to defect in these kinds of circumstances as she did to enter the punishment agreement in the first place?

There is one construction of the contractarian theory of punishment on which revocation of the agreement would be incredibly difficult, if not altogether impossible. As I noted in the previous chapter, the two-tier contractarian test for the moral permissibility of a punishment practice (requiring that an individual both (i) has good reason to agree to the practice, and (ii) agrees to the practice) eventually boils down to whether or not the individual could previously expect to benefit under the practice. In Chapter 3 I began my outline of the contractarian theory by considering a higher threshold than this—namely, one according to which an individual must be guaranteed to benefit under the practice. This assessment would not be made from either an ex ante or ex post position, but would instead be a hypothetical claim that even if the worst-case scenario occurred, the individual would still be better in a world with the practice, than without it. So, when a punishment practice guaranteed a benefit to an individual, it would do so perennially. This is important, because an individual’s attempts to revoke any punishment agreement would do nothing to change that fact. The very thing that made a punishment practice
morally permissible in the first place (a guaranteed benefit) would continue to endure in spite of any attempted revocation. As such, an individual could be taken as constantly continuing to hypothetically agree to the practice from one moment to the next.

Unfortunately, things are not quite so simple when it comes to our lower threshold of expected benefit. This assessment is, by necessity, carried out ex ante and built upon certain assumptions about how future events will play out. While a guaranteed benefit is just that—guaranteed—an expected benefit is not. While an individual may previously have expected to benefit from a punishment practice, it may very well turn out that the practice will not actually benefit her. Unforeseen factors might intervene, and it may become true that if the individual had known how events would transpire, she would not have expected to benefit from the practice and therefore would not have had good reason to agree to its implementation. Given this, it seems that an individual can—by her words or actions—in principle revoke the hypothetical agreement that underpins a punishment practice. So the question remains: What reason does she have to refrain from doing this?

Why individuals have reason to adhere to their agreements—especially where defection may now benefit them more—is a general concern with the wider theory of contractarianism, and not a problem specific to a contractarian theory of punishment. It is, essentially, a restatement of the egoistic “Foole” in Hobbes’s *Leviathan*. While Hobbes’s initial contractarian argument may show that we have good reason to enter certain agreements with our neighbours, it seems to fail to provide a reason why self-interested rationality should also lead individuals to adhere to these bargains over time. As David Gauthier notes, if “it be ever so advantageous to make an agreement, may it not then be even more advantageous to violate the agreement made?
And if advantageous, then is it not rational?” The Foole notes that while mutual co-operation is preferable to mutual defection, defection by one party combined with the co-operation of another is even more advantageous—at least for the defector.

Hobbes attempts to respond to the Foole, arguing that:

“He… that breaketh his Covenant, and consequently declareth that he thinks he may with reason do so, cannot be received into any Society, that unite themselves for Peace and Defence… nor when he is received, be retained in it.”

Hobbes argues that an individual who enters into an agreement only to defect will not last long in a society of individuals who intend for their agreements to be binding. In this way, the Foole will soon find himself ostracised and no longer party to the benefits of co-operation. His making-and-breaking of agreements will ultimately lead to him receiving a lesser utility than making-and-adhering to those same agreements.

This may be true—but it relies on one very important assumption: that is, that the Foole’s agreement-breaking will be detected by those around him. If it will, then Hobbes’s argument stands. The Foole will be ostracised and left in a worse position than if he had kept his agreement. The same is not true if the Foole can break his agreement undetected however. As Gauthier notes:

“Is not the Foole’s ultimate argument that the truly prudent person, the fully rational utility-maximiser, must seek to appear trustworthy, an upholder of his agreements? For then he will not be excluded from the co-operative arrangements of his fellows, but will be welcomed as a partner,

---

222 Gauthier, p. 161.
while he awaits opportunities to benefit at their expense—and, preferably, without their knowledge, so that he may retain the guise of constraint and trustworthiness.\textsuperscript{224}

If the Foole is able to do this, then he will reap the benefits of defection while avoiding the detriments of detection.

But Gauthier proposes a solution to this. He concedes that individuals can, on occasion, successfully deceive their neighbour regarding their intentions. In this way, the intentions of others are not entirely transparent. That being said, individuals’ dispositions are not entirely opaque either—and it is this observation that leads Gauthier to plump for the intermediary concept of “translucency.”\textsuperscript{225} Translucency acknowledges that individuals’ dispositions are sometimes apparent to others, and sometimes not. Whether or not an individual will recognise their neighbour’s disposition is therefore \textit{probabilistic}. Whenever we plan to defect, we run a risk that our intentions will be discovered. This will subsequently motivate other parties to the agreement to also defect, and therefore lead to the worst outcome for all concerned. In light of this risk, Gauthier moves the discussion away from a consideration of what it is rational for an individual to do \textit{in each given circumstance} and instead argues that it is in each of our best interests to develop an ongoing \textit{disposition to cooperate}.

In doing this, Gauthier draws a distinction between ‘straight-forward maximisers’ and ‘constrained maximisers’. A straight-forward maximiser (SM) is an individual who “seeks to maximise his utility given the strategies of those with whom he interacts.”\textsuperscript{226} An SM will, in each circumstance, consider the anticipated behaviour of his neighbour (the individual with whom he is entering an agreement) and calculate how best to maximise his utility given that information.

\textsuperscript{224} Gauthier, p. 161 (emphasis added).
\textsuperscript{225} Gauthier, p. 174.
\textsuperscript{226} Gauthier, p. 167.
The Foole is a straight-forward maximiser, an individual who—in light of the co-operation of his neighbours—will see defection as his most rational course of action.

A constrained maximiser (CM) is instead someone who “has a conditional disposition to base her actions on a *joint strategy*, without considering whether some individual strategy would yield her greater expected utility.”\(^{227}\) Given the observation that mutual co-operation (peace) will yield greater benefits than mutual defection (the condition of nature), the CM adopts a disposition to co-operate. This means that the CM will co-operate whenever she encounters a co-operative agent—regardless of what benefit she might have gained from instead defecting and exploiting the co-operator.

How does this policy of constrained maximisation apply to the specific case of punishment? Agreements regarding particular punishment practices comprise a subset of a large web of agreements (both hypothetical and actual) that form the fabric of our moral and political society. We each fare better by being party to these co-operative arrangements. The translucency of our intentions means that by defecting—or even intending to defect—we run the risk of losing such benefits, and so should adopt a disposition to cooperate wherever possible. This cooperation should extend to all agreements to which we are a party, including those agreements regarding punishment. In this way, Gauthier’s approach provides good reason for an individual to not withdraw his agreement to a particular punishment practice, even where he now finds himself subject to its sanctions. This does not mean that an offender must receive his punishment with open arms. He may—per Hobbes—protest. But this will do nothing to affect the moral permissibility of enforcing the original agreement against the offender.

Unfortunately, several problems remain. First, despite translucency, it seems that there are conceivable scenarios in which the probability of detection is sufficiently low to justify

\(^{227}\) Gauthier, p. 167.
defection. Second, the replies we have so far given to the Foole are all dependent on the fact that our cooperative opportunities will be diminished if we are found to be defecting or intending to defect. Ultimately, this is only contingently true. While unlikely, it is at least conceivable that an offender may defect from a punishment agreement without creating any adverse impact on his future opportunities for bargain-making with others—even if her defection is detected.

It seems, then, that the most rational person will be the Constrained Maximiser who makes certain Exceptions (hereafter referred to as a CME). This is an individual who mostly cooperates, but who defects in those rare circumstances where she is sufficiently assured that either (i) her defection will not be detected, or (ii) her defection—if detected—will not adversely affect her future opportunities for bargain-making with others. The CME personifies the final—and most well-developed—form of the Foole objection. In order to respond to such an individual, we need to point to some further reason that she has to not renege on her agreement, even where no negative consequences are likely to be attached to a particular case of defection.

One potential rejoinder might be found in Michael Bratman’s argument that people’s lives go better insofar as they possess ‘diachronically confirmed planning agency’.

Put simply, Bratman argues that it is better for a person to keep their assurances than not, as the latter involves the serious fragmentation of an individual’s practical reasoning. This stems from the fact that defection involves a conflict between having a reason to adhere to one’s sincerely held long-term plan, and having reason to pragmatically pursue what brings the greatest benefit in the moment. But putting aside the relative merits of Bratman’s argument, it is unclear how much assistance it can render in dealing with this final version of the Foole. It seems as though the CME does maintain diachronic confirmed planning agency when she defects, because taking the

---

228 Bratman, p. 664.
229 Bratman, p. 664.
opportunity to defect in certain favourable circumstances was always a part of her plan. In this way, her practical reasoning is not fragmented. Her standards are deeply consistent. In light of this, we must look elsewhere for a satisfying response to the CME.

An alternative response might be found in the doctrine of fair play that we have already outlined. George Klosko identifies the very problem of the Foole, noting that:\textsuperscript{230}

If we assume that cooperation is perceived as burdensome by those cooperating and the benefits will be provided to a wide range of people whether or not they cooperate, it is in people’s interest to enjoy the benefit without cooperating, that is, to be “free-riders.”

The principle of fairness, however, provides a response by holding that “if others are willing to cooperate and the benefits in question are provided, [the Foole] is also obligated to cooperate.”\textsuperscript{231} As Rawls so succinctly states the maxim: “We are not to gain from the cooperative labours of others without doing our fair share.”\textsuperscript{232} The doctrine of fair play does provide something of a solution to those who—like the Foole—seek to obtain all of the benefits of a cooperative venture without suffering any of the detriments. But, as with the notion of “obligation” discussed earlier, we will encounter difficulty in trying to make sense of certain concepts—like fairness—on a purely contractarian approach. The self-interested rational individual will not have reason to do something merely because it is ‘fair’ to do so. Some other self-regarding reason must be provided.

As it happens, there might be a much simpler solution to this problem—a solution which stems from the very basic notion of what it means to agree to something in the first place. In

\begin{footnotesize}
\begin{enumerate}
\item Klosko, p. 354.
\item Rawls (1971), p.112.
\end{enumerate}
\end{footnotesize}
order for an agreement to make a punishment practice morally permissible, it must *bind* the parties: it must be an agreement to punish—and be punished—from here on out. And a binding agreement is just that: binding. It cannot be broken. A punishment agreement which is capable of revocation at a party’s pleasure is not an agreement at all. It is conceptually incoherent. If an individual *is* now capable of revoking the state’s ability to punish her, then no true punishment agreement ever existed between the individual and the state. Consider again the CME. Either at some prior time the CME did hypothetically agree to the punishment practice, or she did not. If she did not agree, then it was never morally permissible to punish her, and (due to the unanimity requirement) the implementation of the punishment practice was always without moral basis. If she did agree, then this agreement—by its very nature—is incapable of revocation, and remains in effect. The offender may profess to rescind the agreement, and couple this with protestations and attempted evasions of punishment such as those outlined earlier. But this will not matter. The contractarian argument I have presented here is a way of ensuring that the infliction of intentional harm on an offender is morally permissible. This will not require that an offender willingly submits to her punishment. It merely requires that there is a continuing punishment agreement in effect. The irrevocable nature of punishment agreements means that, once hypothetically agreed to, these agreements continue to exist—and provide moral justification—in perpetuity.

**Conclusion on Problem Cases**

In this chapter, I have attempted to describe a number of methods for addressing certain problem cases that may arise in the implementation of our contractarian theory of punishment. As noted, a
central concern for this theory is ensuring unanimous hypothetical agreement to punishment practices via universal expected benefit. The strategies described above are provided in order to ensure that we are not too quick to judge which individuals may or may not provide problems for the theory by failing to expect a benefit, and therefore gaining an effective veto against the implementation of a punishment practice over the wider society. While the approaches outlined above are not intended to guarantee unanimous hypothetical agreement from every individual for every punishment practice, they are intended to clarify the exact scope of individuals who can be brought ‘into the fray’ with the resources currently available within the theory. This clarification also has the added bonus of circumscribing the widest ambit of morally permissible punishment practices for implementation.
Chapter 7: A Solution to the Problem of Punishment

In *The Problem of Punishment*, David Boonin argues that punishment is never morally justified. He does this by systematically demonstrating the unsatisfactory nature of a range of traditional theories of punishment. In the preceding chapters I have presented a new theory of punishment—a theory that, if successful, allows us to once again hold that (at least some forms of) punishment are morally permissible.

In this chapter I will revisit the objections raised by Boonin against both retributivist and consequentialist theories of punishment, and demonstrate that contractarianism is either immune—or capable of replying—to each of these. First, however, I will discuss one specific concern with the potential practical implementation of the contractarian theory.

**A Concern with the Practical Implementation of the Contractarian Theory**

The theory of punishment I have presented in this thesis is not intended merely as a theoretical justification for the moral permissibility of certain punishment practices. It is also meant to provide a practical blueprint which a legislature might use as the basis for an entire institution of punishment—an institution that includes a comprehensive and effective range of punishment practices. At this final juncture, it is worth considering whether what I have outlined here might meet such a purpose.

In Chapter 5, I made reference to a concern with the practicability of the contractarian theory of punishment, stating that:233

---

233 See above at page 113.
The most concerning problem cases for the contractarian theory are those individuals who are not expected to benefit under a particular punishment practice, even where the vast majority of the population will. Such individuals might include those who will go on to offend at an alarmingly high rate, those who stand to gain little from the promised benefit of deterrence, or those simply with value systems far outside the norm. It is these individuals who will deny a practice unanimity, and therefore prevent it from attaining the status of moral permissibility on the pure contractarian approach.

I recommended that when dealing with these individuals, it was useful to split them into two separate classes based upon their epistemic status: cases of non-expected benefit that we know of, and cases of non-expected benefit that we do not know of. I dealt with the first of these categories in Chapter 6, suggesting a number of alternative strategies that might be used to demonstrate how purported expected non-benefiters can expect to benefit from the punishment practice in question. I now turn to consider the second category of non-benefiters: those of whom we are ignorant.

When using the contractarian theory to construct a punishment institution, we should aim for punishment practices that promise an expected benefit to as many individuals as possible—that is, that are maximally rights-respecting. This means that known cases of non-expected benefit will not be tolerated. Put simply, if we know that a certain identifiable individual (or class of individuals) does not expect to benefit under a particular punishment practice (even after applying the battery of strategies outlined in Chapter 6) then it will be morally impermissible to implement that practice.

Things become less clear when it comes to cases of non-expected benefit that we do not know of, or could never know of. Understandably, we want those who are implementing punishment practices to go some way toward investigating to what extent a practice will benefit
the individuals to whom it applies. But we cannot expect this investigation to go so far as making a thorough inquiry into the potential utilities and disutilities of a punishment practice for *every* individual to whom it applies. To require this would be to create a practical burden almost as onerous as requiring the express consent of every individual. Further, the more in-depth our investigation is, the more it will cost us—contributing to the collateral cost of a punishment practice, and potentially leading us to accept certain non-maximal practices. Some inference and estimation of benefit must therefore be permitted, provided that due diligence has been done, and the government has made suitable effort to obtain information about the likely effects of a punishment practice on a general population.

The practicable implementation of our ideal contractarian theory should therefore tolerate the possibility that there may be *some* unknown (or unknowable) individuals who will not expect to receive a benefit under a punishment practice. This is not to say that our ignorance of these individuals justifies subjecting them to the practice. Rather, it is an acknowledgement that while the practice remains unjustified, we at least did the best that we could given our limited epistemic position. Put another way, the implementation of a morally permissible punishment practice does not require the government to have positive and indubitable proof that a punishment practice is expected to benefit every citizen. Instead, there must simply be an absence of proven cases of *non*-expected benefit (after certain sufficient investigation). So long as this is the case, the punishment practice will remain morally permissible when implemented.

This move presupposes what Gideon Rosen refers to as the ‘Ignorance Thesis’: 234

---

Whenever an agent acts from ignorance, whether factual or moral, he is culpable for the act only if he is culpable for the ignorance from which he acts.

Specifically, the case we are here concerned with falls into the category described by Alex Guerrero as that in which “a person is ignorant because, though she has thought about the issue, she has come to have false beliefs about $F$ (she believes that non-$F$ when in fact, $F$).” The Ignorance Thesis is underpinned by what Guerrero refers to as moral epistemic contextualism—that is, the claim that:

How much one is morally required to do from an epistemic point of view with regard to investigating some proposition $p$ varies depending on the moral context—on what actions one’s belief in $p$ (or absence of belief in $p$) will license or be used to justify, morally, in some particular context.

Put another way, “what is required of us from an epistemic point of view increases as what is at stake from a moral point of view increases: our epistemic obligations grow as the relevant context becomes more morally serious.” This seems right. In fact, moral epistemic contextualism may prove incredibly useful to the contractarian theory of punishment—going some way toward describing the extent of investigation a government must engage in when considering the implementation of a particular punishment practice. Put another way, it indicates that the investigative burden required of the state will increase along with the severity of the

---

235 That is, the case in which (1) a government investigates for the existence of non-expected-benefiters under a potential punishment practice PP, (2) finds no evidence of non-expected-benefiters under PP, and therefore (3) implements PP, but (4) where non-expected-benefiters do actually exist under PP.
237 Guerrero, p. 69.
238 Guerrero, p. 70.
punishment practice being implemented. An investigation into the existence of non-expected benefactors under a practice that prescribes execution as a punishment will, for example, bring with it much greater epistemic obligations than an investigation into the existence of non-expected benefactors under a practice that merely prescribes a small fine.

In fact, Guerrero builds upon moral epistemic contextualism to argue that the moral stakes involved in the death of an individual are so high as to never tolerate any kind of epistemic risk. He frames this as the ‘Don’t Know, Don’t Kill’ (DKDK) principle:239

If someone knows that she doesn’t know whether a living organism has significant moral status or not, it is morally blameworthy for her to kill or to have it killed, unless she believes that there is something of substantial moral significance compelling her to do so.

According to Guerrero, DKDK is a principle “that requires caution in the face of a particular kind of moral uncertainty.”240 Further, he claims that those who violate DKDK “act recklessly, and are morally culpable for so acting.”241 Stated more formally, the claim is that an agent is morally blameworthy for an act where that agent:

1. Is non-culpably ignorant of a relevant moral fact,
2. Knows that it is ignorant,
3. Acts from ignorance, and
4. Kills a living organism
5. Without believing that anything of substantial moral significance compels her to do so.

239 Guerrero, pp. 78-79.
240 Guerrero, p. 83.
241 Guerrero, p. 83.
242 Guerrero, p. 96.
DKDK is directed specifically at cases of ambiguity regarding the moral status of a living organism, but it seems as though it can also apply to cases of the state acting so as to end the life of a citizen via execution. Explicitly modified, the claim would be as follows: the state is morally blameworthy for carrying out an execution where it:

1. Is non-culpably ignorant of a relevant moral fact,
2. Knows that it is ignorant,
3. Acts from ignorance, and
4. Executes a citizen
5. Without believing that anything of substantial moral significance compels it to do so.

On this formulation, the DKDK principle is particularly concerning for the contractarian theory of punishment, as it means that the state will still be morally blameworthy for any cases of non-expected benefit, even where the state has carried out sufficient investigation into the existence of potential non-expected benefitters (that is, where it is non-culpably ignorant).

There are three ways in which the contractarian might respond to the DKDK principle: The first of these is to merely deny the principle altogether. Earlier, however, I claimed that the foundation of DKDK—that is, epistemic contextualism—not only seems plausible, but may prove very useful for the practical implementation of a contractarian system of punishment. In that case, denying the DKDK principle will instead require denying that epistemic contextualism implies DKDK. That project falls far outside of the scope of this thesis, however.

Our second strategy is to look for an ‘out’ via the principle itself. The most promising way of doing this is via the final condition, which states that moral an agent might be exculpated from moral blameworthiness for breaching DKDK where “there is something of substantial
moral significance compelling it to do so.” What might carry sufficient significance? One potential candidate is the preservation of the lives and safety of a large number of people. We might, for example, argue that it is morally permissible to risk a severe detriment (that is, execution) to a small number of people if it can be shown that the wider punishment practice brings an overwhelming benefit to the wider community (say, by drastically reducing the murder rate). This, however, will be completely unacceptable to the contractarian. The very basis of the theory of punishment outlined in the preceding chapters has been the (somewhat controversial) claim that in order for a punishment practice to be morally permissible, it must be agreed to by everyone—and that this in turn requires that every individual can expect to benefit under the practice in question. To now claim that a certain number of non-benefiting individuals will be tolerated, provided that the vast majority of people do benefit, is to allow the theory to devolve into mere consequentialism and undo all of our previous work.

What options remain, then? It seems that the only justification of “substantial moral significance” tolerated under the contractarian theory of punishment would be positive proof of the universal benefit—and therefore unanimous agreement—of all affected parties. If we had this, however, the issue of DKDK would not even arise in the first instance, as the state would no longer fall prey to the issue of not knowing how its citizens would be affected.

The third, much simpler, strategy for responding to the DKDK principle presents itself when we realise that (unlike the Ignorance Thesis from which it is borne) the principle is referring specifically to ignorance of moral facts. In the case of the state using contractarianism to justify the death penalty, then, the first condition (that the state is non-culpably ignorant of a relevant moral fact) is not met. The state is, in fact, well aware of all moral facts regarding under what circumstances the death penalty would be morally justified (these are the moral facts
provided by the theory I developed in chapters 2, 3 and 4: specifically, that in order for the implementation of a punishment practice involving the death penalty to be morally permissible, every individual must expect to receive a benefit under that punishment practice). What the state may struggle to know is whether these conditions are actually met for each individual, as this would require a deep, onerous (and, in some cases, perhaps impossible) inquiry into the personal valuations—and expected future actions—of every one of its citizens. Put simply, the potential epistemic ignorance suffered by the state in applying the contractarian theory is non-moral. While there may be some argument to be made regarding moral culpability stemming from the non-culpable ignorance of non-moral facts, this is not what the DKDK thesis addresses. As such, the contractarian theory manages to escape Guerrero’s argument.

Where does this leave us? Previously, I stated that “the practicable implementation of our ideal contractarian theory should… tolerate the possibility that there may be some unknown (or unknowable) individuals who will not expect to receive a benefit under a punishment practice.” I stand by this claim. It will be morally excusable for a state to implement a punishment practice with something less than positive proof that all individuals can expect to benefit under that practice. Wilful ignorance does not exculpate, however, and the state will have to go some way towards satisfying itself that universal expected benefit exists. Precisely what the state’s epistemic obligations are will differ from punishment to punishment and will, as per epistemic contextualism, increase according to the severity of harm caused by the punishment in question. If, during the mandated investigation, cases of non-expected benefit are uncovered, these must be addressed using the strategies outlined in Chapter 6. If none of these strategies succeed—and the non-expected benefit withstands—the punishment practice will fail to bring an

243 See above at page 157.
expected benefit to all, thereby causing it to lack unanimous agreement and entailing the moral impermissibility of the practice in question.

**How Contractarianism Avoids Concerns with Traditional Theories of Punishment**

Under my contractarian theory of punishment, the moral justification for a punishment practice is derived from the unanimous hypothetical agreement of those to whom it applies. I have argued that such agreement will exist where every one of those individuals can expect to benefit from the implementation of that practice. These theoretical underpinnings of the contractarian approach mean that is able to circumvent many of the objections raised against other theories of punishment.

Take, for example, Desert-Based Retributivism. As we saw, the biggest concern with this theory is that moral desert and legal guilt do not always coincide. This is of no concern to the contractarian theory, however, as desert does not even enter into our considerations. Instead, the contractarian theory is concerned simply with the bare moral permissibility of punishment—permissibility that will arise wherever (1) an individual can expect to benefit from the implementation of a punishment practice, and (2) goes on to commit the crime that is covered by this punishment practice. Likewise, the contractarian solution makes no mention of goods—or an unequal distribution of such goods—and therefore does not fall prey to the same kinds of problems as Fairness-Based Retributivism.

The theory I have presented also manages to avoid the issues that arose with Forfeiture-Based Retributivism, including: (1) identifying which rights are forfeited when an individual commits a crime, (2) explaining the rights-equivalency problem, (3) specifying for how long a
right is forfeited, (4) preventing individuals from carrying out private retaliation against offenders whose rights have been forfeited, and (5) solving the problem of how removing something to which an offender has no right (because that right has been forfeited) can amount to harm, and therefore punishment. The contractarian theory of punishment manages to avoid all of these issues by simply removing rights from the equation. In fact, the contractarian theory doubly insures itself against particularly unpleasant outcomes—like private retribution—as individuals are highly unlikely to agree to the implementation of a punishment practice that could put them at risk of such occurrences.

Nevertheless, the contractarian approach is not entirely above reproach. As stated above, the normative force of the theory stems from a foundation of *expectation of benefit*—and the source of this benefit is the deterrent effect of the punishment practice in question. If a punishment practice $X$ effectively deters others from carrying out an action that a person $P$ disvalues, then $P$ receives some kind of benefit from the implementation of $X$. But this reliance on deterrence means that the contractarian theory of punishment shares a genetic similarity with that of consequentialism—a theory to which Boonin raises a significant number of objections. While contractarianism and consequentialism diverge significantly in their treatment of deterrence (and, as a result, their conclusions on specific cases), their foundational similarity means that some of Boonin’s objections to consequentialism may—with appropriate modifications—be levelled successfully against contractarianism.

One of the most common arguments against theories of punishment rooted in deterrence is the Punishing the Innocent Objection. This is the concern that, in certain circumstances, greater deterrence may be achieved by punishing an innocent party. In Chapter 1, we considered a case in which law enforcement could attain a gross increase in utility by punishing an innocent
individual in order to avoid rioting that would result in much greater harm to others. Consider, now, a modified case in which the punishment of an innocent individual would specifically lead to an increase in deterrence: Suppose that there has been a bank robbery for which no perpetrator can be found. As the weeks go by, and law enforcement fail to come up with any leads, more and more individuals begin to believe that bank robbery is a relatively risk-free endeavour, thus increasing their otherwise latent desire to carry out such a crime. In response, law enforcement officers severely and publicly punish a single innocent individual whom the public are made to believe is guilty of the robbery. Suppose further that this display completely deters all other individuals from even entertaining the notion of robbing a bank.

A theory focussed solely on a wholesale maximisation of deterrence would, it seem, endorse such actions. Is this similarly a problem for the contractarian theory of punishment? I would argue that it is not. First, sometimes we will have good reason to agree to a punishment practice by which the innocent are routinely punished. This might sound counter-intuitive, but consider an example to demonstrate this: Suppose that everyone has an opportunity to agree to the implementation of a government-run ‘Strange Lottery’. The features of this lottery game are as follows: Each week, every member of the lottery has a 99% chance of winning. Every win is independent of other wins, so it is possible for every participant to win in the same week. Upon winning, the successful entrant receives $1000. If the entrant loses, however, he or she is forced to pay a penalty payment of $10 to the government. It seems that everyone will have good reason to agree to enter such a lottery. While every entrant runs a 1% risk of a $10 loss each week, this is far outweighed by the promise of a 99% chance of receiving $1000. In purely probabilistic terms, in any given week the lottery has an expected return of $989.90 ([0.99 x 1000] + (0.01 x -10)) for each individual.

---

244 At least, everyone who places some value on financial resources.
Consider that we now modify this example slightly. Suppose that instead of winners receiving $1000 in cash, they receive state-provided protection of their person and property valued at $1000. Suppose, further, that participants will be given this protection both when they win and when they lose. Will individuals still have good reason to agree to such an arrangement? It seems so. Every entrant now has a 100% chance of receiving a $1000 benefit each week – a benefit for which they must simply shoulder a $10 penalty approximately every one hundred weeks.

Suppose, then, that we make one final adjustment. As well as receiving the $10 penalty, the winner will also receive a (false) declaration that he has committed some arbitrary offence against the person or property of another—and that the penalty is to be treated as punishment for that offence. Would individuals still have reason to agree to this scheme? I contend that they would. Provided that there are no further ramifications of the allegation that an offence has been committed (that is, assuming the $10 fine is the only harm suffered by a lottery loser) each individual will still fare far better under this system than they would without it. Although it involves the punishment of the innocent, everyone—even the punished—are better off. This seems to suggest that individuals will, in certain circumstances, plausibly have good reason to agree to a system of punishment that punishes innocent people.

Thus, it is false to claim that—in principle—individuals will never have good reason to agree to a system of punishment that routinely punishes the innocent. For this reason, it makes sense that contractarianism might end up endorsing such a practice. The theory I have prevented claims that a punishment practice will be morally permissible where it receives unanimous

245 That is, suppose that without this protection each entrant would—on average—suffer approximately $52,000 of harm per year to his person and property.
246 Of course, in practice, there are many flow-on effects of being punished. These would need to be considered when calculating the full value of Detrimental Punishment for a case like this.
agreement from all to whom it applies. Further, it claims that such agreement will exist where all individuals, at some prior time, could have expected to benefit under that practice when compared with the status quo. Given this, a punishment scheme such as the Strange Lottery is almost guaranteed a contractarian endorsement.

The range of scenarios in which this will be the case is severely limited, however. And the reason for this is simple: deterrence. The ‘punishment’ meted out by the Strange Lottery is—by stipulation—in no way expected to provide any kind of deterrence. Instead, this example merely talks of the state providing ‘protection’ of life and property in the abstract. In almost all real-world cases of punishment, however, such protection comes *by way of* deterrence. As noted in the previous chapter, however, a punishment practice will only have a deterrent effect where the commission of a crime *increases an offender’s likelihood of receiving that punishment*. A punishment that is administered without regard for innocence or guilt will be no deterrent at all, and will thus provide no benefit—and therefore no reason—for individuals to agree to its implementation. The receipt of punishment would be a mere lottery, with the commission of a crime doing nothing to increase one’s chances of receiving a sanction.

Return, then, to the case of the bank robbery. The contractarian theory would only endorse the punishment of an innocent individual in this case where there was unanimous agreement from all to whom that practice applied (including the unfortunate soul who is punished). As stated above, such agreement will only exist where everyone, at some prior time, could have expected to benefit under that practice when compared with the status quo. Such benefit—if it existed—would derive entirely from the amount of deterrence being purchased. But, for the reasons just mentioned above, punishing the innocent can provide no such deterrence. No individual could expect to benefit from a punishment practice that fails to provide
deterrence, and thus such a practice would fail to receive the kind of agreement the contractarian agreement requires in order for a punishment to be deemed morally permissible.

In some cases, however, ‘punishing the innocent’ may be less blatant than the deliberate targeting of individuals whom law enforcement knows to be innocent. Most Westernised criminal law systems have ‘beyond reasonable doubt’ as their standard of evidence. The intention behind such a high threshold is to minimise the number of innocent individuals who are wrongfully convicted. The corollary of this, however, is that some guilty parties walk free. For the reasons stated above, the fact that these offenders go unpunished works against the deterrent efficacy of punishment. By reducing the standard of proof (say, to the civil standard of ‘on the balance of probabilities’) more of these guilty offenders would be successfully convicted, thereby increasing deterrence levels. At the same time, however, the number of convicted innocent individuals would presumably also increase. Would such a move be justified under the contractarian theory of punishment? The short answer is “yes,” provided that there was a universal expectation of benefit for all to whom the new standard of evidence applied. It must not only bring an expected benefit to the unpunished populace, but also to those individuals who find themselves punished—including those who are in fact innocent. Such a state of affairs is conceivable (consider, again, the example of murder reduction discussed in Chapter 3), and in such a circumstance the contractarian theory would properly endorse such an evidential modification.

At the opposite end of the spectrum to the Punishing the Innocent Objection is the Not Punishing the Guilty Objection—a concern that arises wherever the harm caused by punishing the offender exceeds the benefit gained in deterrence. As noted in Chapter 1, this might occur where the offender is a pillar of the community, or responsible for the provision of some kind of
unique and incredibly valuable service, without which the community will suffer immensely. Once again, however, the contractarian theory has an answer to this concern. The theory—at bottom—is about eliciting the agreement of individuals by protecting what they value by deterring those behaviours that they disvalue. A punishment practice which failed to punish certain individuals for certain crimes would fail to provide effective deterrence, and thus would provide little reason for others to agree to it. Few citizens would agree to a punishment practice which gave free license to certain individuals to break the law without punishment simply because of a position they hold or a role they serve within the society.

Variations of the Punishing the Innocent and Not Punishing the Guilty Objections also arise for retributivists theory—though these stem from a different issue to that suffered by deterrence-based theories. As discussed in Chapter 1, the origin of these objections for retributivists arises from a mismatch of moral desert and legal guilt. Individuals such as the vehement racist may deserve to be punished, while in fact being legally innocent (thus leading to cases of the innocent being punished), while individuals such as the emissions law-breaker may not deserve to be punished despite being legally guilty (thus leading to cases of the guilty not being punished). The contractarian theory avoids this conundrum entirely by doing away with any notion of desert. An individual is punished not because she deserves to be punished, but because she has performed an action that—by unanimous hypothetical agreement—is to be punished when performed. This approach means that there will always be a perfect coincidence between when (i) it is morally permissible to punish an individual, and (ii) that individual is legally guilty of a crime.

A further concern with theories rooted in deterrence is the Disproportionate Punishment Objection. This arises from the observation that, in most circumstances, a harsher penalty will be
a more effective source of deterrence. While a $500 fine may provide a decent amount of deterrence for the crime of bicycle theft, prescribing the death penalty for such an offence will garner even greater deterrence. It seems, however, that in most circumstances we will be reluctant to endorse the implementation of such severe penalties. The contractarian theory of punishment accommodates this intuition. While harsher penalties may increase the level of deterrence, they will also increase the disutility of Detrimental Punishment (that is, the amount of detriment that an individual will suffer if he or she becomes subject to the punishment in question). In order for a punishment practice to be justifiable, it must remain the case that even those individuals who expect to be punished can expect to be better off under the practice than they would have been under the status quo. If the punishment is too excessive, this condition will not be met, and there will be no resultant hypothetical agreement to provide a moral basis for the practice.

Deterrence-based theories also often find themselves falling prey to the No Excuses Objection, when mitigating factors—many of which are often, and rightfully, taken into account in legal sentencing—serve no role in dictating the type and severity of punishment an individual receives. If deterrence of a particular kind of crime is all that matters, then it should be of no concern whether an offender had good reason to commit a crime, or was compelled, or in a dissociative state at the time of commission. The contractarian theory, however, is able to take such factors into account. While the deterrence of disvalued actions is a prime motivator for people agreeing to the implementation of a punishment practice, other factors will come into play too. What is of central concern is whether or not an individual can expect to benefit under the practice, and when making these calculations, we must inevitably consider the harm an individual will suffer if she herself commits the crime. In doing so, individuals will
understandably shy away from any punishment practice that sanctions without regard for robust mitigating circumstances.

Finally, deterrence-based theories often find themselves subject to the Treating People as a Means Objection. This is because the punishment of an individual is used purely as a way of reducing further incidences of a particular offence. A punishee is a mere means to the end of crime reduction. But while consequentialist theories are vulnerable to such an allegation, the contractarian theory is not. By hypothetically agreeing to a punishment practice, an individual necessarily agrees to be punished for reasons of deterrence should they commit the offence in question. I contend that the presence of such agreement means that an individual is no longer treated as a means if (or when) she is used for the purposes of deterrence.

Consider an analogy: when I order a coffee from a barista, there is a sense in which I am using him for a particular goal—namely, my attainment of caffeination. It does not make sense to claim that he is being used as a means to an end, however. This is because my utilisation of him for the purposes of attaining caffeination is with his agreement. He consents to be used in such a way, namely because he receives a benefit in the form of gainful employment. He wants customers to order coffee, as without such patronage his café would fail and he would no longer be employed. By providing him with my custom, then, I am in fact carrying out an action that attains both the ends of myself and the barista. The same is true in the case of punishment. Every person who hypothetically agrees to a punishment practice does so because he or she expects to benefit from the deterrence which stems from that practice. When we punish an individual, we do so with the intention of attaining that end for all parties to the practice—including the end of the party who is now being punished. Some concern might be raised by the fact that—unlike the barista—each individual’s agreement to a punishment practice is only hypothetical rather than
express. But it is important to remember that this hypothetical agreement is derived from what the individual has good reason to agree to, and that this assessment is in turn made on the basis of what the individual values—that is, on the basis of her ends. It is a truism, then, that any individual who is treated according to a hypothetical agreement of this kind will necessarily not be treated as a means.

In this way, then, the contractarian theory of punishment is able to uniquely escape those problems that ordinarily plague approaches rooted in deterrence. Its ability to do this stems from the fact that the theory is, at bottom, focussed not merely on wholesale deterrence, but rather on attaining an expected benefit for all to whom it applies. At the same time, its conceptual framework sees it also circumvent many issues face by other non-deterrence based theories of punishment.

Conclusion

Consent brings with it a strong normative force. In this thesis, I have used that force to explain how we might morally justify intentionally harming individuals in the name of punishment. The theory I have provided will not necessarily endorse the precise system of punishment with which we are all familiar. In fact, the punishment practices prescribed by a contractarian theory may look very different to anything we currently employ. At bottom, the foundation of this theory is deterrence, and the pursuit of said deterrence may demand the implementation of punishment practices and standards of evidence that are highly counterintuitive. Nevertheless, it is important to remember the point from which this argument began. If Boonin’s contention is true, no current theory provides a justification for the moral permissibility of punishment. The theory I have put
forward here is an attempt to provide a coherent alternative—a theory of punishment that escapes
Boonin’s critique, and provides an encouraging solution to the problem of punishment.
Bibliography


The Innocence Project, “The Cases: DNA Exoneree Profiles” <www.innocenceproject.org/cases-false-imprisonment>


The Crimes Act 1961 (New Zealand)

The Sentencing Act 2002 (New Zealand)

*Chappell & Co Ltd v Nestle Co Ltd* [1960] AC 87


*R v Brown* [1993] UKHL 19

*R v Cato* [1976] 1 WLR 110 CA

*R v Johnson* (1986) CA