Utilitas (2019), **31**, 246–261
doi:10.1017/S0953820819000104

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Bentham on Temptation and Deterrence

Steven Sverdlik

Southern Methodist University
Corresponding author. sverdlik@smu.edu

(Received 14 October 2018; revised 22 March 2019; accepted 22 March 2019)

Abstract

In *Introduction* Bentham considers a difficulty. If the immediate aim of punishment is to deter agents considering breaking the law, then the severity of the threat of punishment must increase if they are strongly tempted to offend. But it seems intuitively that some people who were strongly tempted to offend should be punished leniently. Bentham argues in response that all potential offenders capable of being deterred must be deterred. He makes three mistakes. (i) It is possible that it would produce the most happiness at t2 to punish an offender who could have been deterred at t1, but was not. (ii) The Principle of Utility might condemn the threats that would be needed to deter all potential offenders who can be deterred. (iii) Given the dispositions to reoffend of some strongly tempted offenders, their punishments should be relatively lenient. There is more room for leniency in Bentham's theory than he realized.

Jeremy Bentham was a pioneer in the philosophy of punishment – this is generally known. And the basic ideas of his theory are also generally known: punishment should be designed to promote the greatest happiness in society, and this would mainly be achieved by deterring potential criminals from committing offences. Nonetheless, it is striking how few of his detailed analyses and arguments in this area have been carefully assessed.[1] This article will focus on some important arguments he makes in applying the principle of utility to the activity of deterrence.

I examine some passages in *An Introduction to the Principles of Morals and Legislation* (1789). This book was originally intended to be an introduction to a model penal code, designed on utilitarian lines (1, 4). It is the best-known statement of Bentham's philosophy of punishment.[2] In chapter 14, 'Of the Proportion between

[1]One outstanding exception is H. L. A. Hart, *Punishment and Responsibility* (Oxford, 1968), pp. 17–21. This criticizes the 'rationale of excuses' in Chapter 13 of Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, ed. J. H. Burns and H. L. A. Hart (Oxford, 1996), pp. 160–2. Numbers in the text refer to pages in this edition.

[2]The passages in Bentham's *Introduction* that I focus on have parallels in Bentham's *The Rationale of Punishment*. Much of *Rationale* was written in the mid-1770s, although parts of it were clearly written decades later. It was not published in English until 1830. See Jeremy Bentham, *The Rationale of Punishment*, ed. James McHugh (Amherst, NY, 2009), p. 13. *Rationale* seems to contain earlier versions of some chapters in *Introduction*. Book I, Chapter 6 of *Rationale* closely parallels Chapter 14 of *Introduction*; Book I,

Punishments and Offences', Bentham presents rules which are meant to determine the correct level of severity of the punishments for different crimes. But he acknowledges an apparent difficulty concerning Rule 1.

> Rule 1: The value of the punishment must not be less in any case than what is sufficient to outweigh that of the profit of the offence. (166)

The difficulty is this: if the immediate aim of punishment is to deter agents considering breaking the law, then the severity of the threat of punishment must increase if they are strongly tempted to offend, since in that case the 'profit of the offence' will be great. But in some cases it seems intuitively that people who were strongly tempted to offend and did so should be punished leniently, not severely. Bentham's response to the difficulty is nuanced. He grants that in some cases leniency is called for. However, he insists that leniency must never go so far as to result in a punishment that fails to deter all strongly tempted agents who can be deterred. Bentham's critics have argued that his conclusion requires objectionably severe punishments for some such offenders.

I will argue that Bentham's response involves three mistakes about how the principle of utility governs the practice of legally authorized deterrence. Bentham himself therefore was mistaken in insisting that utilitarianism requires punishment that will deter all strongly tempted potential criminals who can be deterred. It can allow more leniency than he realized.

The issue here is not as narrow as it might seem. First of all, Bentham's argument provides us with a nice specimen of his dialectical skills in defending his philosophical theory. I believe that in *Introduction* Bentham acknowledges only one specific difficulty with the theory, and he then proceeds to respond to it. This is the difficulty about temptation and deterrence. Second, as I said, Bentham's position has troubled a number of commentators over the years, including the most distinguished of them, A. C. Ewing and H. L. A. Hart.[3] Finally, the misgivings of these commentators are echoed in contemporary non-consequentialist philosophy of punishment.[4] In showing that utilitarianism can allow more leniency than Bentham realized I want to suggest that it contains theoretical resources incompletely understood by all concerned: Bentham himself, as well as his critics.

I. Rule 1 and Bentham's psychological hedonism

Chapter 14 of *Introduction* states 13 rules that determine the morally correct degree of severity of punishments for legal offences. Bentham clearly thinks that these rules

Chapter 4 of *Rationale* parallels Chapter 13 of *Introduction*. See n. 24 below, where I argue that one passage in *Introduction* seems to be assuming an argument only made explicitly in *Rationale*. I cite *Rationale* by book and chapter, as well as by pages in *The Works of Jeremy Bentham*, ed. John Bowring, vol. 1 (Edinburgh, 1843).

[3]A. C. Ewing, *The Morality of Punishment* (London, 1929), pp. 52–4; H. L. A. Hart in Bentham, *Introduction*, p. cv. See also James Fitzjames Stephen, *Liberty, Equality, Fraternity* 2nd edn., ed. R. J. White (Cambridge, 1967), pp. 152–4; Elie Halevy, *The Growth of Philosophic Radicalism*, trans. Mary Morris (New York, 1928), pp. 69–70. F. W. Maitland, 'The Relation of Punishment to Temptation', *Mind* 5 (1880), pp. 259–64, largely agrees with Bentham. Anthony Draper, 'Punishment, Proportionality, and the Economic Analysis of Crime', *Journal of Bentham Studies* 11 (2009), pp. 1–32, is the only previous study that tries to situate Bentham's arguments in Chapter 14 of *Introduction* in his larger theory.

[4]Michael Moore, *Placing Blame* (Oxford, 1997), p. 29.

https://doi.org/10.1017/S0953820819000104 Published online by Cambridge University Press

jointly articulate, via more specific considerations, what the principle of utility directs legislators to do when they construct codes that specify the punishments for legal offences. It is after he states Rule 1 that the difficulty about strongly tempted offenders is presented and answered. This important rule implies claims about which psychological factors influence the decision to commit a crime (or 'offence'), and thus draws on the psychological hedonism that Bentham presents elsewhere in *Introduction*. Such a psychological theory entails claims about how to deter potential offenders. However, Rule 1 is not correctly formulated, given Bentham's psychological hedonism. Since the psychological details are unimportant for my purposes, I proceed quickly to a better formulation of it. I then state Bentham's psychological hedonism.

Here again is Rule 1.

> Rule 1: The value of the punishment must not be less in any case than what is sufficient to outweigh that of the profit of the offence. (166)

This rule is implicitly claiming that three basic psychological factors influence the decision to commit an offence. 'Value' is a concept Bentham applies to both pleasures and pains; 'profit' refers only to pleasures. As he originally explains it, value includes (i) the *quantity* of a pleasure or pain (further distinguished into its intensity and duration); its (ii) *certainty* or probability of occurring; as well as (iii) its '*proximity*', that is, proximity in time (38–9).[5] These are the three basic factors in Bentham's psychological hedonism. However, all three need to be understood in Rule 1 in subjective terms. For example, the quantity of a pleasure or pain there refers to the amount of pleasure or pain that a potential offender *believes* (or 'expects') to experience if she commits a crime. Likewise, the operative notion of certainty refers to the degree of certainty that *she has* about the occurrence of such pleasure and pain.[6] What Bentham means to convey by 'proximity' is that the proximity or remoteness in time that an agent believes her pleasures and pains will have can modify their impact on her decisions (169–70). A pleasure or pain that an agent believes she will experience in the distant future can have less impact on her decision than one that she believes will occur closer in time, even if both involve the same subjective quantity and probability. We can call this discounting factor the 'subjective proximity' of future pleasures and pains.

The error in Bentham's statement of Rule 1 is the following. The rule mistakenly focuses on the difference that an agent believes will come about between the amount of pleasure and the amount of pain for her that will be produced by one of her options, namely, the offence. It should instead focus on the differences in the *net* amount of pleasure the agent believes will come about for her in *all* of her options.

These points can be illustrated as follows. Simplifying somewhat we can say that Bentham's Rule 1 entails that if an agent S believes that committing a crime C will produce 5 units of pain for her and 4 units of pleasure for her then she will not commit C. Here the pain for her presumably 'outweighs' the pleasure for her. In net terms, and representing units of pain with negative numbers, she believes she will experience −1 unit of pleasure net. This cannot be what Bentham is assuming. Suppose that S believes that the only alternative to doing C is obeying the law in some way, O. S might believe that if she does O then she will experience 6 units of pain and 4 units

[5]There Bentham speaks of 'propinquity' and 'remoteness'. 'Proximity' occurs at *Introduction*, pp. 169–70.

[6]This important shift in meaning is clear in his explanation of the meaning of 'profit', but the same point obviously applies to 'value'. See *Introduction*, pp. 166, note c ('expectation of profit'), 167 ('apparent profit').

https://doi.org/10.1017/S0953820819000104 Published online by Cambridge University Press

of pleasure. The net amount of pleasure she believes she will experience by doing O is −2 units. This is worse for her than the −1 unit she believes she will experience by doing C. Bentham would surely say that in that case S will choose to commit C.[7]

If we utilize Bentham's terminology then Rule 1 is best stated as follows. (The redundant term 'net' is included to make clear that an agent may believe that some of her options will produce both pleasure and pain for her.)

> Best Version of Rule 1: The punishment of an offence must be such that a potential offender believes that the net value for her of the offence is less than the net value for her of obedience to the law.

Here is a paraphrase:

> The punishment for an offence must threaten potential criminals who are considering committing it in such a way that the following is true: they believe that, when the net amount of pleasure for them of committing it is discounted by its subjective probability and subjective proximity, there will be a greater net amount of pleasure for them in obedience to the law, when that is also discounted by its subjective probability and subjective proximity.

We can take the difficulty about temptation and deterrence to concern this version of Rule 1.

I will now state Bentham's psychological hedonism. It is an oddity of *Introduction* that Bentham nowhere explicitly states this proposition. The reader has to assemble all of the factors that Bentham states are relevant to human decision-making from separate passages, including Rule 1.[8] Keep in mind that 'value' includes both subjective probability and proximity. The redundant term 'net' is again included.

> Bentham's Psychological Hedonism: All human beings always choose to perform the action, among those which they believe it is possible for them to perform, which they believe has the greatest net value for themselves.

II. The objection and Bentham's reply

After stating Rule 1 Bentham acknowledges that it has often been objected to, sometimes by 'authors of great merit and great name', 'on account of its seeming harshness'.[9]

[7]Bentham seems to recognize this point at *Introduction*, p. 162, paragraph 11.

[8]The fundamental claims of Bentham's psychological hedonism are presented at *Introduction*, pp. 96–100. In this material, Bentham does not mention the factor of 'proximity'. See *Introduction*, pp. 169–70. Bentham's later 'A Table of the Springs of Action', *Deontology*, ed. Amnon Goldworth (Oxford, 1983), pp. 79–115, largely follows *Introduction*, although it seems to abandon psychological hedonism at p. 100, and tends to neglect the roles of certainty and proximity.

[9]*Introduction*, pp. 166, note c, 167. Bentham does not identify these authors. It is clear that William Eden (1745–1814) is his main target. See William Eden, *Principles of Penal Law* (London, 1771), pp. 7–9. The reference to 'authors' is thus puzzling. Eden's book went through four editions by 1775, but all were published anonymously. However, Bentham knew that Eden was the author: *The Correspondence of Jeremy Bentham*, ed. Timothy Sprigge, vol. 2 (London, 1968), pp. 100, 114. Eden was a prominent politician in the late 1770s, when Bentham was writing *Introduction*. Eden's eminence may have made Bentham reluctant to criticize him by name. Two other authors are mentioned in Radzinowicz's authoritative work in

https://doi.org/10.1017/S0953820819000104 Published online by Cambridge University Press

The allegedly problematic implication of Rule 1 and Bentham's first response go as follows.

> True it is, that the stronger the temptation, the less conclusive is the indication which the act of delinquency affords of the depravity of the offender's disposition. So far then as the absence of any aggravation, arising from extraordinary depravity of disposition, may operate, or at the utmost, so far as the presence of a ground of extenuation, resulting from the innocence or beneficence of the offender's disposition can operate, the strength of the temptation may operate in abatement of the demand for punishment. But it can never operate so far as to indicate the propriety of making the punishment ineffectual, which it is sure to be when brought below the level of the apparent profit of the offence. (167)[10]

Bentham gives no examples of the cases he has in mind; neither the ones calling for some reduction in severity, nor those calling for none. Nor does he define 'temptation' explicitly. We can start to fill in these gaps.

Bentham does define the 'strength' of the temptation that an agent experiences with regard to a possible action of hers. In fact, he gives two different definitions of it. The first is the ratio between the amount of pleasure that an agent expects to receive from her action and the amount of pain she expects to receive from it, both discounted by their subjective probability and proximity (138).[11] The second is simply the amount of pleasure that she expects to receive from the action, presumably discounted by its subjective probability and proximity (166–7). On either account Bentham conceives of almost any agent who performs an action as tempted to perform it. The only possible exception would be cases where an agent expects all of her options to produce only pain for her. In any case, this expansive concept of temptation helps to explain why he does not think that it always operates to mitigate the punishment for performing the action it favours.

H. L. A. Hart presented a pair of cases that, he suggested, put Bentham's position about deterring tempted offenders to the test. They occur in Hart's discussion of Rule 1, and he finds Bentham's position troubling.

> [A] starving man who steals a loaf would, other things being equal, be punished more severely than a rich man stealing something for which he cared little. (cv)[12]

relation to *Introduction*, pp. 166–7: Blackstone and Paley. But they basically agree with Bentham. Leon Radzinowicz, *A History of English Criminal Law and its Administration from 1750* (London, 1948), p. 384. Cf. Draper, 'Punishment', pp. 21–31.

[10]Bentham seems still to have accepted this argument later in his life. Jeremy Bentham, 'Specimen of a Penal Code', in *Works*, vol. 1, pp. 164–8, at 166–7.

[11]Bentham does not explicitly mention there subjective proximity, but it is included in the Best Version of Rule 1, and is explicitly mentioned in Rule 8 (*Introduction*, p. 170).

[12]The example of theft of food by a starving person, and the claim that leniency is appropriate, are common. Thomas Aquinas, *Summa Theologica* (1274), II-II, 66, 7; Thomas Hobbes, *Leviathan*, ed. C. B. Macpherson (Harmondsworth, 1968), ch. 27, marginal heading, 'Total Excuses', p. 346; William Blackstone, *Commentaries on the Laws of England* (Chicago, 1979), vol. 4, p. 15; Maitland, 'Relation', p. 262; Ewing, *Morality*, p. 52. Bentham himself discusses such an example. *Introduction*, p. 140. Cf. the slightly different example in *Rationale*, I 6, p. 400.

Let us set aside for now the rich man. It is true that if Bentham was claiming that Rule 1 establishes that punishing the starving man harshly – say, by four years in prison – is morally required, then his utilitarianism seems objectionable. We will consider whether it does establish this.

III. The relevance of Chapter 13

In understanding Bentham's argument and its weaknesses, it is helpful to put it in context. Two features of its context are important for my purposes: the other rules presented in Chapter 14, and the discussion in Chapter 13.

Let us begin with Chapter 13, which is titled, 'Cases Unmeet for Punishment'. In this famous chapter Bentham describes four categories of action where utilitarianism opposes the punishment of agents who perform them. I believe that Bentham's argument in Chapter 14 is based on the assumption that the actions he is describing do not fall into any of these four categories. I will describe them.

First there is Bentham's category of 'cases in which punishment is groundless' (159–60). These are acts which, though they may usually cause 'mischief' (that is, pain), do not do so in some cases, or cause mischief that is 'outweighed' by the pleasure also produced. Nowadays in the philosophy of punishment it is said that the acts falling into this category should be treated as 'justified' legally.

Second, there is the category of acts where 'punishment must be inefficacious' (160–2). The reason it must be inefficacious, Bentham says, is that in these cases the agent could not be deterred from committing her offence. Punishing such offenders does not bring about the good of crime prevention, and causes pain to the offender. Therefore, utilitarianism says that such acts of punishment are themselves wrong. Some examples that Bentham gives of cases where punishment must be inefficacious are offenders who are insane, ignorant of certain facts, or subject to coercion (161–2). Nowadays in the philosophy of punishment it is said that the agents performing actions in such psychological conditions should be granted 'legal excuses'.[13]

The third category consists of acts where 'punishment is unprofitable' (163–4). These acts cause mischief and are wrong, and the agent should have no legal excuse (since she is deterrable). However, the pain of any punishment 'would be greater than what it prevented' (159). Bentham's main concern here is an issue that we would now describe as 'criminalization', that is, whether the criminal law is a cost-effective way of decreasing the incidence of 'pernicious' forms of behaviour (287). *Introduction* has two passages where Bentham considers such cases (163–4; 287–91; cf. Rule 12, 171). He mentions two forms of 'self-regarding' behaviour that he thinks the criminal law is largely powerless to prevent in a cost-effective way: drunkenness and fornication. The problem in criminalizing such behaviour is that obtaining evidence of them would require violations of privacy and 'tearing the bonds of sympathy asunder' (290).[14]

[13]Hart uses this term, and it is now in general use: Hart, *Punishment*, pp. 13–14. Hart confined the domain of legal excuses to various kinds of psychological limitations of offenders, and this, too, is now generally accepted. *Punishment*, p. 14.

[14]He also mentions four types of case where a type of act should be criminalized, but 'occasional circumstances' render it unprofitable to punish one or more offenders. One example he gives is where there are so many offenders that punishment of all of them will not produce the most happiness. *Introduction*, pp. 163–4. The issue discussed in sect. VI is structural, not occasional.

https://doi.org/10.1017/S0953820819000104 Published online by Cambridge University Press

The fourth category of cases unmeet for punishment consists of acts where punishment is needless (164). This is actually a sub-category of cases where punishment is unprofitable.

When we reach Chapter 14 Bentham seems to be assuming that these four categories of action have been set aside. Therefore, the actions under consideration do cause mischief, so their punishment is not groundless. Also, the punishment of the agents is not inefficacious, so the agents are deterrable, and their punishment (if properly proportioned) will produce the most happiness. And, finally, we are dealing with acts that should be criminalized.

We should pause here to explore a few further points.

First, we should recognize that some of the actions that utilitarianism says should be treated as legally justified or legally excused are thefts of bread, and, in fact, thefts of bread by starving people. Bentham may not have realized the following point, but it is certainly consistent with his theory to say this: one type of situation, or one psychological condition, abstractly described, may operate now as a justification, now as an excuse, now as a mitigation (reducing but not eliminating punishment), and now have no influence on the severity of punishment. In other words, different tokens of the type have different moral statuses, and should therefore have different legal statuses.[15] We can take the abstract type to be 'stealing bread to stay alive'.

Consider how a token of this type could be morally right according to utilitarianism. If stealing a loaf of bread from, say, a supermarket were the only way that a starving person could save his life, then a utilitarian would say such an act is generally right, not wrong, since it would produce more happiness than its alternatives. Bentham could thus claim that utilitarianism will sometimes say a starving man who steals bread ought to be treated as having been legally justified in so acting. This means that the objection made by Hart and others to the utilitarian approach to punishing strongly tempted offenders is at least partly unfounded. Utilitarianism will say that some starving men who steal bread should not be punished at all – much less, punished severely.[16]

Now consider why a utilitarian would say that some tokens of the type 'stealing bread to stay alive' should be treated as legally excused. A starving man acts from a very powerful desire, and may be undeterrable.[17] Therefore, a utilitarian can say that in some cases a starving man acts morally wrongly, but excusably. For example, a man dying of starvation may steal a loaf of bread from a supermarket, despite the

[15]Cf. Hart, *Punishment*, p. 16; George Fletcher, 'The Individualization of Excusing Conditions', *Southern California Law Review* 47 (1974), pp. 1269–88, at 1274 (with regard to necessity). Bentham may have been thinking along the same lines. *Introduction*, p. 162, paragraph 10, seems to assert that some cases of mistake about the existence of circumstances that would provide legal justification for acting should be granted excuses.

[16]Bentham notes that the profit of an offence is not always 'proportioned to the mischief' (*Introduction*, pp. 168–9, note k). But he does not see that this point, which he discusses in relation to Rule 2, is relevant to the argument he makes about deterring people subject to strong temptation. In some cases, the profit of stealing bread is so great that instances of it should not be an offence at all.

[17]Given Bentham's conceptions of when an act should be treated as legally justified and when it should be treated as legally excused, it would be possible for one act to be such that it should be both legally justified and excused. This is because someone who is undeterrable could perform an act which produces the most happiness. When these conditions hold it would be better to say that it should only be treated as legally justified. In the rest of this paragraph I am discussing acts which should not be treated as justified, but should be excused.

fact that it had a rack of highly discounted older baked goods in the back of the store. The starving man may have had enough money to buy a discounted loaf, but was so crazed with hunger that he grabbed the first loaf he saw, and would not have paused to consider the existence of other options, no matter what he believed the punishment for his theft was. The utilitarian will say that punishment of this man would be inefficacious, and therefore wrong. The utilitarian will therefore say that this starving man also should not be punished at all – much less, punished severely.

Hence, to focus our inquiry, we need to construct a certain version of Hart's case. We can assume that utilitarianism will generally favour criminalizing stealing bread. We need a case where a starving man acts wrongly in stealing it, and, according to utilitarianism, should have no legal excuse for so acting. This presumably gives us a case where utilitarianism requires us to punish him, as Hart's example does not necessarily do. So consider this scenario. Some of its details will be important later on.

> *The Nearly Starving Man*. A poor man was nearly starving. He had repeatedly sought help for two days in getting food, but failed to procure it. His hunger grew intense and he strongly desired to steal a loaf of bread that he saw in a supermarket. He did steal it. However, the supermarket had a rack of highly discounted older baked goods in the back of the store. The starving man had enough money to buy a discounted loaf. If he had believed that there was a punishment of at least four years in prison for stealing this bread, he would have investigated whether there was another way to procure bread besides stealing, and he would have learned of the discounted food rack. He did not believe that the punishment was this severe. Had he learned of the discounted bread, he would have bought a loaf instead of stealing one.

I will say that when we are considering acts that should generally be criminalized, and which should not be treated as legally justified or excused, we are dealing with 'eligible' offences. Their agents are 'eligible' offenders. The Nearly Starving Man, I will assume, has committed an eligible offence, and he is an eligible offender. We will proceed to consider whether utilitarianism would favour punishing him with four years in prison.

IV. The other rules in Chapter 14

We can now turn to the other feature of the context of Bentham's argument about temptation and deterrence, Chapter 14 itself. He gives the argument there, and states it immediately after Rule 1. Bentham thus seems to regard that rule as deciding the matter.[18] But there are other rules mentioned in that chapter. This strongly suggests that no matter how carefully Rule 1 is stated, and its scope limited to eligible offences and offenders, it has a *ceteris paribus* character.

The other rules do present important considerations that have to be weighed in a utilitarian system of legal punishment. Two other rules are particularly important.

> Rule 2: The greater the mischief of the offence, the greater is the expense, which it may be worth while to be at, in the way of punishment. (168)

[18]Bentham takes Rules 7, 8 and 9 to be closely related to Rule 1. Rules 7 and 8 are entailed by the Best Version of Rule 1. Rule 9 brings in a distinct consideration, namely, the likelihood that someone who commits one offence had committed others, and was not punished for them. *Introduction*, p. 170.

https://doi.org/10.1017/S0953820819000104 Published online by Cambridge University Press

> Rule 5: The punishment ought in no case to be more than what is necessary to bring it into conformity with the rules here given. (169)

Rule 2 states that punishments may be more severe if they prevent more harmful crimes, and Rule 5 requires that punishments be frugal, presumably with regard to the pain they cause the criminal and the expense they impose on taxpayers (cf. 179). These two rules implicitly and roughly articulate the basic moral considerations relevant to the setting of severity levels mandated by the principle of utility: the moral benefit of crime reduction, and the moral costs of the criminal's pain and of legal administration.[19] Set in this context, Rule 1 is by no means decisive.[20]

We might say that the principle of utility requires lawmakers and judges to apply a certain sort of moral cost-benefit analysis to the practice of legal punishment. We can also say, of course, that it requires lawmakers and judges to maximize the amount of happiness in society.

V. Bentham's first mistake

I now argue that, given his own assumptions, Bentham made three mistakes in his argument that punishments must be severe enough to deter all strongly tempted potential eligible offenders. In this section, I describe his first mistake.

I will assume henceforth that the discussion only concerns eligible potential offenders. And I will assume that the principle of utility, with its distinctive moral cost-benefit analysis, governs the design of the criminal law. Finally, I will mention something that Bentham surely understood, namely, that a system seeking to deter potential offenders has to be analysed as a structure with two temporal stages. The first stage is the time at which the criminal law in effect makes threats to potential criminals; then the hope is to induce them to choose to obey the law. The second stage is the time at which punishments are imposed. The imposition of these punishments can be described as carrying out the threats made at the previous time. To save words in discussing this temporal structure, I will say that the time of the first stage is 't1'. The second stage I will say is 't2'.

Bentham elaborates on his response to the critics of Rule 1, quoted before, in an important passage. In defending this rule against the charge that it entails that unduly severe punishments are sometimes morally required Bentham concedes, we saw, that a reduction in severity may sometimes be allowed for some of those who are strongly tempted to break the law. Still, he continues, the reduction can never be so great as to bring it below the level needed to deter them. He says this:

[19]Many of the other rules can be seen as implications of these two rules. Rules 3 and 4 can each be seen as an implication of Rule 2. Rule 12 can be seen as an implication of Rule 5. Rules 6, 10 and 11 can be seen as implications of the conjunction of Rules 2 and 5. (Rule 13 is second-order, requiring simplicity in the set of legislated rules.) *Introduction*, pp. 168–71.

[20]When we understand the normative force of Rules 2 and 5, Rule 1 comes to seem superfluous. Rule 1 focuses on a single causal channel of crime reduction, deterrence via a threat antecedent to an offence. However, punishment can reduce crime in other ways, as Bentham knew. For example, punishment can prevent crimes by disabling a criminal after an offence (*Introduction*, pp. 181–2). All of the channels of crime reduction should be governed by the utilitarian cost-benefit analysis, which is roughly articulated by Rules 2 and 5. This normative analysis can be stated at a level of abstraction in which deterrence antecedent to an offence has no special role.

> The partial benevolence which should prevail for the reduction of it below this level, would counteract as well those purposes which such a motive would actually have in view, as those more extensive purposes which benevolence ought to have in view: it would be cruelty not only to the public, but to the very persons in whose behalf it pleads: in its effects, I mean, however opposite in its intentions. Cruelty to the public, that is cruelty to the innocent, by suffering them, for want of an adequate protection, to lie exposed to the mischief of the offence: cruelty even to the offender himself, by punishing him to no purpose, and without the chance of compassing that beneficial end, by which alone the introduction of the evil of punishment is to be justified. (166–7)[21]

This passage occurs in Chapter 14, and parallels what he writes in the preceding chapter on cases where punishment 'must be inefficacious'. Indeed, he uses the same word, and gives a reference to that chapter (166–7).[22] In Chapter 14 Bentham is saying that punishing a person who *non-excusably* breaks the law because the threatened punishment was too lenient is 'inefficacious' and therefore wrong according to the principle of utility: it will cause pain, but not produce the most happiness. However, the kind of case we are here considering is fundamentally different, given Bentham's assumptions, from the ones he discussed in Chapter 13.

Bentham argues in the passage just quoted that utilitarianism tells us that it would be wrong to punish an eligible offender at t2 who had been strongly tempted to commit an eligible offence at t1, and could have been deterred by the threat of a harsher punishment, but was not. Punishing her at t2, he says, would be inefficacious: it would not produce the most happiness.

Here is why he is mistaken. At t2 the offence has already occurred, and by hypothesis, cannot then be prevented. The facts that the offence at t1 was mischievous and morally wrong, as well as the fact it could have then been prevented, have no relevance as such to the question of what action *at t2* will produce the most happiness. The 'forward-looking' orientation of utilitarianism tells us to consider whether any offences *after t2* can then be prevented. It is often said that carrying out threats at t2 by imposing punishments maintains the 'credibility' of the system of threats and punishments, and this can be seen to promote happiness. Maintaining the credibility of the system can be decomposed in Bentham's terms into two main causal channels: the effect of the punishment on the offender himself, and its effect on other agents who may be considering whether to break the law.[23] Someone who was deterrable at t1, but was not deterred, will usually be deterrable at t2. If so, she will be responsive to the actual imposition of punishment, and this could cause some change in her propensity to commit crimes.[24] It

[21]Bentham recurrently contrasts 'partial' or 'confined' and 'extensive' or 'enlarged' benevolence, the latter being the more likely to be objectively correct. *Introduction*, pp. 117–18, 128, 135.

[22]This discussion in Chapter 14 lays out the utilitarian argument against inefficacious punishments more explicitly than the well-known passage in Chapter 13.

[23]When punishment actually imposed diminishes an offender's tendency to break the law Bentham speaks of its 'reforming' effect (*Introduction*, p. 180–1). When it diminishes the tendency of other agents to break the law he speaks of its serving as an example, or its exemplarity (*Introduction*, pp. 178–9). Bentham also recognizes that some punishments simply 'disable' an offender from committing crimes (*Introduction*, pp. 181–2). Cf. n. 20 above.

[24]Bentham seems to deny this when he says that such punishment is cruelty 'to the offender himself, by punishing him to no purpose'. He does not spell out his reasoning, but it is presumably given at *Rationale*, I 6, p. 399. There he says, 'If … a man, having reaped the profit of the crime, and undergone the punishment,

could do this by causing her to believe that the amount of punishment she will suffer is greater than she had thought; that the probability she will undergo it is greater than she had thought; that her future pains are more important to her than had been true previously; or some combination of these. The punishment of an offender at t2 may also serve as an example to other deterrable agents, so that they make the same sorts of changes to their beliefs or values, or some combination of them.

In short: if the Principle of Utility is true then it is possible that it is right to punish at t2 an eligible offender like the Nearly Starving Man for an eligible offence at t1 – even though he could have been deterred at t1 from committing it by the threat of a harsher punishment for it, but was not.

VI. Bentham's second mistake

My criticism of Bentham may seem inadequate, since it is asserting that a punishment of an offender at t2 can be right even if a greater threat to her at t1 would have deterred her. But, you may say, Bentham uses Rule 1 to argue that the threat at t1 should have been severe enough to deter her. Indeed he does.[25] Given the Best Version of Rule 1 his claim is this: the law must threaten every eligible potential offender with a punishment severe enough to ensure that she believes that the net value for her of the offence is less than the net value for her of obedience to the law.

Bentham makes another mistake here. He fails to understand how the moral considerations implied by Rules 2 and 5 modify the force of Rule 1 as it applies to threats, as well as to the punishments that carry them out. The principle of utility does not require the law to threaten every eligible potential offender with a punishment severe enough to deter her.

To see the mistake we will first continue to explore the point made in the last section: it can produce the most happiness to punish an eligible offender who was not deterred. The next question to pose is this: how severely should such an offender be punished? When Rules 2 and 5 are used to answer this question we can see that they will not always mandate severity levels that deter every eligible potential offender. This important conclusion will then be used to understand the correct severity levels of threats. We begin, though, with acts of punishment.

In invoking Rules 2 and 5 to analyse punishment severity I now describe the important empirical possibility of diminishing deterrent power of increases in the severity of

finds the former more than equivalent to the latter, he will go on offending'. Bentham is presumably picturing an eligible offender with a standing desire to commit a given offence, whose strength does not change after punishment. Such an offender is also being pictured as having subjective probability and proximity values that do not change after punishment. Bentham is arguing that if such a person believes, after undergoing the punishment for it, that he is better off than he was before committing it, then he will go on committing it. In that case, punishing him (at the same severity level) is wasted pain, at least with regard to preventing his offending. However, there are other types of offender: punishing them after failing to deter them can deter them in future. For example, after punishment an offender may change his belief about the probability of being punished. If so, the expected net value for him *after t2* of offending again may be less than the expected net value for him of obeying the law, even if the expected net value for him *at t1* of committing the offence was greater than the expected net value for him of obeying the law.

[25]Speaking of an error made by Anglo-Saxon laws he writes that it is also made by any legal system 'wheresoever the punishment is fixed while the profit of delinquency is indefinite: or, to speak more precisely, where the punishment is limited to such a mark, that the profit may reach beyond it' (*Introduction*, p. 167). This passage follows his statement of Rule 1.

https://doi.org/10.1017/S0953820819000104 Published online by Cambridge University Press

punishment. In fact, there is evidence that this pattern commonly occurs.[26] It can (and does) occur when some eligible potential offenders are more difficult to deter, so that the costs of deterring them grow, while the benefits of crime reduction stay constant or decline. In terms of the offence we are considering the pattern is this: at a certain point in severity, it ceases to increase total happiness to prevent thefts of a loaf of bread by making offenders suffer more.

To see this I will suppose that the act of stealing a loaf of bread is criminalized, and that three levels of punitive severity are legally permitted for the theft of one loaf of bread. Suppose further that each instance of a theft of a loaf causes on average 2 units more of victim pain than thief pleasure, or a net loss to society of 2 units of pain (as compared to not stealing). Now consider how different levels of severity might affect the incidence of thefts of loaves of bread. The facts assumed are summarized afterwards in a table.

> Level 1: thieves who steal one loaf are punished with 3 units of pain, and there is one more unit of pain due to the expense of administration. Each act of punishment will prevent 3 thefts. One such act of punishment thus produces a benefit of 6 units, and a net benefit (after subtracting the pain of the criminal and the expense of administration) of 2 units.
>
> Level 2: thieves who steal one loaf are punished with 6 units of pain, and there is one more unit of expense. Each act of punishment will prevent 5 thefts, yielding a benefit of 10, and a net benefit of 3.
>
> Level 3: thieves who steal one loaf are punished with 9 units of pain, and there is one more unit of expense. Each act of punishment will prevent 6 thefts, producing a benefit of 12 and a net benefit of 2.

	Total Cost of Punishment	Benefits from Deterrence	Net Effect on Total Happiness
Level 1	−4	+6	+2
Level 2	−7	+10	+3
Level 3	−10	+12	+2

Utilitarianism says that the severity level of acts of punishment for stealing one loaf that is most cost-effective morally is Level 2, 6 units of pain for the criminal.

Now, if the punishment is set at 6 units of pain, and this is the amount that is threatened at t1, then an eligible potential offender who is considering whether to steal a loaf of bread, but would only be deterred by the threat of a punishment of 7 units of pain or more, will not be deterred. If he is not deterred he may nonetheless be apprehended and punished for stealing the bread. If so, and given our assumptions, utilitarianism says that the severity of his punishment should be 6 units of pain.

My argument can be seen as making assumptions about the known effects of stealing a loaf of bread, and possible punishments of it. But another version of the argument works if we suppose that the decision to punish is made under uncertainty, as

[26]For a clear discussion, see National Research Council, *The Growth of Incarceration in the United States*, ed. J. Travis, B. Western and S. Redburn (Washington, 2014), pp. 138–40.

https://doi.org/10.1017/S0953820819000104 Published online by Cambridge University Press

Bentham thought it always is (168, 288). We can simply suppose that all the numbers represent expected (social) utility. If so, utilitarianism will say that Level 2 is the most reasonable choice, morally. The same point applies to the arguments about threats, which follow.

How would utilitarianism guide the legislative act of announcing or promulgating levels of punishment severity for stealing a loaf of bread? This can be thought of as the act of issuing a threat to those who are considering whether to steal a loaf of bread, and it will operate at t1. Bentham claims that utilitarianism requires legislators to issue threats that will deter every eligible potential offender. However, Rules 2 and 5 do not always support this claim.

In examining Bentham's claim I assume that we are considering possible new threats that could be made in a society where thefts of bread are already threatened and punished to some degree. Suppose that Level 2 (6 units of pain) is the threat currently in force, and it is always imposed when an eligible theft of bread occurs. By the argument above, this level of actual punishment produces the most happiness, but it does not deter every eligible potential offender. Let us begin by considering whether increasing the threat, while keeping all the other features of the criminal justice system constant, will always succeed in deterring every eligible potential offender. The other features of the system involve the detection of crimes, the trial and sentencing of offenders, as well as the actual punishment imposed. We are assuming that all of these stay constant in order to isolate the effect of threats themselves.

Increasing threats alone seems attractive for a utilitarian since it, unlike actual punishments, seems to have virtually no cost in offender pain or from administration. Still, we must determine how effective as deterrents such increases would be. Would threatening 9 units of pain, say, for stealing a loaf, when only 6 units are imposed, and the other law enforcement factors remain constant, manage to deter every eligible potential offender? Not necessarily. It is very possible that threatening 9 units and imposing 6, with all else constant, will deter more eligible potential offenders than will threatening 6 units and imposing 6. But threatening 9 units and imposing 6 would not guarantee that all eligible potential offenders will be deterred. This sort of bluffing will be detected by those subjected to punishment, and they may well alert others to it. So the extra deterrence achieved over time by the bluffing threat will be limited. And, in any case, some eligible potential offenders may not be deterred by a threat of 9 units of pain that they fully believe will be imposed.

Furthermore, in any actual criminal justice system, an increase in threatened punishment may well have effects on other features of the system. These effects may involve significant costs. For example, a threat to impose a greater punishment may trigger a legal requirement to provide an accused offender with a lawyer, increasing the costs of administration. Once such costs are taken into account it is again possible that at a certain point in threatened severity, it ceases to increase total happiness to prevent thefts of a loaf of bread by *threatening* offenders with more suffering. The utilitarian cost-benefit analysis of acts threatening punishment can mandate that their severity levels be below that needed to deter every eligible potential offender.

The point here is of general application, and does not only apply to the deterrence of agents we would intuitively describe as strongly tempted. It applies, for example, to people who have a weak desire to steal something, and believe that it is very unlikely that they will be caught. But the point does apply to strongly tempted potential offenders like the Nearly Starving Man. Utilitarianism says that a threat to punish him with four years in prison may be too severe, even if that is necessary to deter him.

I believe that Bentham misunderstood the general point because he assumed that the issue of the 'profitability' or cost-effectiveness of punishment (and threats) is largely an issue about criminalization. However, it will commonly arise for actions that should be criminalized, like stealing a loaf of bread.

VII. The third mistake

Bentham erred in claiming that utilitarianism requires that all eligible potential offenders be deterred. So the maximum amount of punishment to be threatened or imposed for stealing a loaf of bread might be low, say, three months in jail. These points apply to the deterrence of any eligible potential offender who is considering such an action. Bentham's third mistake pertains specifically to some strongly tempted offenders. Curiously enough, he misunderstood the implications of some claims that he himself made about their psychological dispositions. When these claims are taken into account, we can see that utilitarianism would favour leniency specifically for offenders like the Nearly Starving Man.

In order to frame the last issue properly, let us return to Hart's worry.

> [A] starving man who steals a loaf would, other things being equal, be punished more severely than a rich man stealing something for which he cared little. (cv)

Perhaps Hart is not thinking that utilitarianism will require punishing the Nearly Starving Man severely in an absolute sense – as my previous stipulation of a four-year term in prison assumed. Hart may only be troubled by the thought that utilitarianism will require that the Nearly Starving Man be punished more severely than the rich man. We will investigate this.

The legal concepts that we need to consider here are the 'extenuating' (or mitigating) and aggravating features of crimes and offenders.[27] The question now is whether utilitarianism requires sentencing the Nearly Starving Man to, say, a three-month jail term, and a rich man who stole an object comparable in value to a loaf of bread to a one-month jail term.

Bentham's thinking on questions of aggravation and extenuation is not presented as systematically in *Introduction* as is that on legal justifications and excuses: there is no chapter or sub-section explicitly devoted to it. It is reasonably clear, though, that he envisioned a legal system in which judges at sentencing would apply some or all of the rules in Chapter 14, or some more specific legal rules, to this issue. In doing this the judge would consider certain characteristics of an offender, and she would have some discretion to determine if these called for increased or diminished severity.[28]

[27]Hart *Punishment*, pp. 14–17, speaks of 'mitigation', which is now the more common term. Bentham speaks of 'extenuation' at *Introduction*, p. 167.

[28]*Introduction*, pp. 69–70, 169, paragraph 15. *Rationale* contains a brief chapter: VI 1, pp. 516–17. Later in his career Bentham developed his thinking on how a utilitarian legal system should be codified, and how judicial officials would be legally authorized to apply these codes to cases, including, presumably, to sentencing. For two opposing interpretations of Bentham's position on the general question of judicial decision-making and discretion, see Gerald Postema, 'Bentham and Dworkin on Positivism and Adjudication', *Social Theory and Practice* 5 (1980), pp. 347–76, at 350–8, and Francesco Ferraro, 'Adjudication and Expectations: Bentham on the Role of Judges', *Utilitas* 25 (2013), pp. 140–60. But Bentham's views in 1780 on these matters were undeveloped. As Hart noted, *Introduction* contains no discussion of constitutional law (Hart, *Introduction*, p. cx; cf. p. 281, n. a). Bentham, 'Specimen', is much more detailed on sentencing than *Introduction*, but still sketchy on the limits of judicial discretion. Furthermore, English sentencing law *c.* 1780 did not provide Bentham with a usable model. It was extremely rigid in

Bentham describes Chapter 11, 'Of Human Dispositions in General', as relevant to sentencing decisions (141–2). The main topic of the chapter is this: given that an offender has performed a specific offence at t1, and that this offence is to be punished, what can we infer about his disposition to offend at t2?[29] Note that this discussion assumes that such a person has offended and, thus, was not deterred.

Bentham concludes the chapter with a discussion of the degree of 'depravity' of an offender that is indicated by a specific offence. He states four rules that summarize what can be inferred about this, given certain psychological facts about the offender at t1. These rules are not moral or normative; they articulate factual presumptions. Bentham explicitly states that an offender's degree of temptation at t1 is one relevant psychological fact (140–1). It is significant that in Bentham's discussion of temptation and deterrence in Chapter 14 he refers back to the section of Chapter 11 that contains these four rules (167). He uses the same terminology ('temptation', 'depravity' and 'disposition') in both places.

In Chapter 11 we find this illustration of its Rule 3, which is stated just before it:

> [I]f a poor man, who is ready to die with hunger, steal a loaf of bread, it is a less explicit sign of depravity, than if a rich man were to commit a theft of the same amount. (140)

Here is Rule 3 of Chapter 11:

> Rule 3: The apparent mischievousness of the act being given, the evidence which it affords of the depravity of a man's disposition is the less conclusive, the stronger the temptation is by which he has been overcome. (140)

In other words, if two offenders at t1 both believe that the amount of mischief or pain that their offence will cause is the same, then the one who was more strongly tempted to commit it is likely to be less depraved. Bentham's reason for saying this apparently is the following: the starving man seems to have had at t1, and is likely to have at t2, stronger restraining (or 'tutelary') motives.[30] Rule 3 can be improved slightly. Bentham notes that if someone yields to a strong temptation it does not follow that she would not also have yielded to a weaker one (141). However, if a temptation gradually grows in strength and a person only yields to it when it becomes strong, she did resist the weaker temptation. So we should revise Rule 3 as follows.

> Rule 3a: The apparent mischievousness of the act being given, the evidence which it affords of the depravity of a man's disposition is the less conclusive, the stronger the temptation is by which he has gradually been overcome.

This rule, like Rule 3, should be understood as being *ceteris paribus*, since Bentham recognizes other relevant facts about an offender's psychology at t1. He argues, for example, that some motives, when favouring acts believed mischievous, also indicate bad dispositions (127f.).

---

theory, at least with regard to felonies. Judges often had no discretion, and were required to sentence convicted criminals to death, even for property crimes. Various mitigating techniques existed, some of dubious legality. John Langbein, *The Origins of Adversary Criminal Trial* (Oxford, 2003), pp. 57–61, 324–5, 334–6.

[29]Draper, 'Punishment', pp. 27–31, discusses this chapter, but largely rests his interpretation on an unpublished manuscript.

[30]For tutelary motives, see *Introduction*, pp. 134–5, 145–6.

https://doi.org/10.1017/S0953820819000104 Published online by Cambridge University Press

Let us consider the implications of Bentham's comparison of the poor starving man and the rich man. He assumes that both offenders have the same beliefs at t1 about the mischievousness of their actions, which is something we can accept for our purposes. And we can stipulate that the poor man is the Nearly Starving Man, who resisted temptation for two days. Finally, we can stipulate that the Nearly Starving Man and the rich man both act from a 'self-regarding' motive. Bentham states that when this motive favours performing an act believed to be mischievous, it indicates a mischievous disposition (127). Given everything Bentham says, or should have said, about the two thieves as we are supposing them to be, and their offences – which includes Rule 3a – the conclusion that follows is this: the rich man is likely to have a more depraved disposition at t2 than the Nearly Starving Man. This means that, other things being equal, the rich man should be punished more severely.[31]

Surprisingly, then, Bentham considers something close to the very comparison that Hart uses to pose a problem for utilitarianism. Bentham reaches the conclusion that Hart thinks is correct, and he supports it with reasoning focused, as utilitarianism is, on the production of happiness at t2. It is also surprising that Hart did not recognize the significance of this passage.

Nevertheless, Bentham goes on to insist in Chapter 14, as we have seen, that the punishment of all eligible offenders must be severe enough to deter them (166–8).[32] Bentham's perceptive discussion of some of the inferences we can make about the dispositions of offenders is ultimately undermined by his mistaken thinking about temptation and deterrence.

We can detach the one from the other, and thus say the following. Given the psychological assumptions Bentham makes about a strongly tempted offender like the Nearly Starving Man, the principle of utility favours punishing him relatively leniently, as compared to offenders who steal something of equal value, but are less strongly tempted to do so.

VIII. Conclusion

Bentham's conclusion about the deterrence of strongly tempted offenders seems to imply that a starving person who steals a loaf of bread should be punished severely. However, his philosophy of punishment actually implies that this type of theft might sometimes be correctly treated as legally justified or excused, and hence not punished at all. Furthermore, this philosophy actually implies that the punishment for any theft of a loaf of bread might have a low upper limit on severity, so that some strongly tempted eligible offenders are not deterred. And, finally, it actually implies that some strongly tempted eligible offenders (like the Nearly Starving Man) ought to be punished less severely than a rich man who steals something comparable. It is striking that neither Bentham nor his critics considered all of these implications in addressing the very issue that Bentham acknowledged as an apparent difficulty. Both Bentham and his critics misunderstood what his theory entails about temptation and deterrence. There is more room for leniency in it than any of them saw.[33]

[31]This disregards the goal of general deterrence, but it is not clear that it would favour punishing the Nearly Starving Man more severely than the rich man.

[32]Cf. *Introduction*, p. 142, which occurs after Rule 3 and its illustration.

[33]I thank three anonymous referees, Dale Miller, Robert Howell, Matt Lockard, Luke Robinson, Charles Curran, Alastair Norcross and, especially, Justin Fisher for helpful comments.

Cite this article: Sverdlik S (2019). Bentham on Temptation and Deterrence. *Utilitas* **31**, 246–261. https://doi.org/10.1017/S0953820819000104