A Moral Aversion Theory of Punishment

Alonzo Fyfe (Draft of 20170116)

Copyright

First Printing: 2017

ISBN <Enter your ISBN>

Table of Contents

Abstract: In this paper, I attempt to provide a motive-consequentialist defense of criminal punishment. Criminal punishment involves intentionally harming another person because that person has broken the law. The fact that it involves intentionally harming others means that it is something that we should avoid doing unless there is some overriding reason to do so. In his book, The Problem of Punishment, *David Boonin surveyed attempts to justify punishment and found them all to be defective. For example, utilitarian theories, which argue that punishment can be justified by its good consequences in terms of utility, imply that we should punish innocent people when it produces good consequences – which is wrong – so it must be rejected. I wish to argue that a motive-consequentialist theory can answer Boonin's objection concerning punishing the innocent. It can also answer the rest of Boonin's objections, thus solving the problem of punishment – contingent upon the verification of certain empirical facts concerning the effects of punishment.*

Condemnation and Punishment

I. Introduction

In his book, *The Problem of Punishment*, David Boonin examined several theories of legal punishment (that is, punishment for breaking the law).[1] He found all of them to be flawed. From this he concluded that criminal punishment is not justified unless it is necessary. He then argued that it is not necessary. Consequently, punishment is not justified.

I argue that Boonin did not give a motive-consequentialist defense of punishment its proper due.

Boonin applied most of his objections to utilitarian theories, arguing that what mattered was not a particular consequentialist theory's account of "the good" but the relationship between the good and the right. (Boonin: p. 80) Through most of this essay, I will phrase my response in terms of how a motive utilitarian would answer Boonin. However, in the end, I will show that this defense will require that we abandon utilitarianism in favor of a pluralistic consequentialism.

A motive utilitarian defense of punishment argues that punishment is something that a person with good motives would support. Insofar as this is a utilitarian theory, motives, in turn, are considered good to the degree that they produce the best consequences in terms of overall utility.

Boonin asserted that motive-utilitarian theories are not significantly different from rule-utilitarian theories. Rule-utilitarian theories say that punishment is justified when it is in accordance with good rules (as opposed to good motives). Good rules, in turn, are the rules that produce the best consequences in terms of utility. In applying his objections against rule utilitarianism to motive utilitarianism, Boonin seems to regard motives as substantially the same as rules.

I will identify several significant differences between rules and motives that make motive-utilitarian theories immune to Boonin's objections to a rule-utilitarian defense of punishment.

I will start by giving a fuller account of the problem of punishment as Boonin defines it, focusing on the rule-utilitarian response to that problem and Boonin's criticisms of that response. I will then look at what motives are and the unique role that punishment plays in modifying motives. From this, I will construct a motive-utilitarian theory of punishment that meets Boonin's criteria

---

[1] Boonin, David (2008). *The Problem of Punishment.* Cambridge University Press. Future references to this work will be indicated by: (Boonin: page numbers).

for a theory of punishment. I will then explain how this motive-utilitarian theory handles Boonin's objection that utilitarian theories justify punishing the innocent. Next, I will examine some of Boonin's other objections to rule-utilitarian theories of punishment and look at what motive-utilitarian theories have to say about those issues.

Ultimately, this account of condemnation and punishment will show itself to be inconsistent with classical utilitarian account of "the good". It will yield a pluralistic form of motive consequentialism instead. These differences will not be relevant through much of the argument that follows, since Boonin's objections are applicable to both versions. However, the need for pluralistic consequentialism will show itself in the end.

II. Boonin's Objections to Punishment

Boonin's objection to punishment generally goes as follows: Punishment involves treating two groups of people differently; those who have broken the law, and those who have not. Specifically, it involves harming members of the first group – and intentionally harming them at that; while, at the same time, it is clearly wrong to intentionally harm members of the second group. Why? (Boonin: p. 28)

To try to answer, "Why?", Boonin went through every major theory of punishment in turn and showed that none of them could remove the presumption against intentionally harming members of the second group. Thus, none of them could justify criminal punishment.

A. Boonin's Anti-Utilitarianism

All utilitarian theories ultimately justify their conclusions based on producing the most utility. The justification for punishment looks to the overall good that punishment will produce and the harms and suffering it helps us to avoid. An institution of punishment, the utilitarian claims, helps to bring about a society in which people generally can live safer and more fulfilling lives. The lives of those punished may be made worse off, but the overall benefits to the community far outweigh those costs.

Imagining a society without prohibitions on theft, rape, and murder tends to bring to mind the benefits that punishment gives us, and the costs that it allows us to avoid - though the fact that

we can imagine certain outcomes or states does not imply that it must necessarily come to pass or that there are no other alternatives.[2]

The utilitarian justification for punishment sounds good on the surface, but Boonin identifies several problems. Chief among these is that the utilitarian defense of punishment also implies that we may – and, in some cases, should - punish the innocent. The only thing that matters on a utilitarian theory is increasing utility. The guilt or the innocence of the person to be punished does not matter, except insofar as it influences the resulting utility. If we can produce utility by punishing an innocent person, we should do so. (Boonin: p. 39ff)

It is not at all difficult to imagine cases in which punishing an innocent person produces good consequences. A common example involves punishing an innocent person to prevent harm to several innocent people.

Imagine a case in which a white police officer has shot and killed a black person. The black person, wielding a heavy pipe, was charging the police officer when the officer fired. After shooting him, the black person's friend took the pipe and ran away, managing to escape. He tossed the pipe away and then told his buddies that his friend (and theirs) had his hands up and had surrendered when the police officer killed him. The police officer's partner witnessed the events and reported them honestly to her superiors.

Let us imagine that this takes places in a context where police officers have been repeatedly caught on video abusing suspects. These include instances of police officers clearly shooting unarmed black men, accompanied by evidence of cops conspiring to cover up the crimes and destroy evidence. Many people assume that this new case is another example of the same type of behavior. Consequently, there are riots – people demanding that this police officer be punished. Refusing to punish this person based on the testimony of his partner will be taken as the police conspiring to protect each other (yet again). If this police officer is not convicted of this crime, the resulting riot will almost certainly get out of hand and several innocent people will die.[3]

---

[2] Boonin, importantly, is not arguing that morality condemns us to live in such an environment. He argues that we have ways of avoiding those costs and harvesting those benefits, or something near enough to them, without punishment. However, this paper will not address his alternative.

[3] This is a modern version of an objection originally raised by H. J. McCloskey, "An Examination of Restricted Utilitarianism." *The Philosophical Review* 66 (1957), pp. 468-9.

Utilitarian theory says, "Of course, convict him. It will produce the best consequences. All that matters is that the harm that this one officer will suffer will be less than the cumulative harms that the riot will inflict on others in the city. This conclusion is absurd, and reason to reject the utilitarian defense of punishment. Whatever it is that justifies punishment, if anything, it is not to be found in acting to produce the most utility. It must exist elsewhere, or nowhere."

B. The Rule Utilitarian Response

As a response to these and other objections to the utilitarian theory of punishment, some philosophers propose a two-tier system of morality.

John Stuart Mill, for example, argued in defense to the objection that we cannot assess the overall utility of telling a lie versus telling the truth each time an opportunity to lie presents itself, answered that we know that – as a rule – lying produced bad consequences. Furthermore, a flourishing society was one that required each person to have confidence that others were telling the truth – not one in which each citizen knew that others were assessing the utility of lying with each opportunity. Consequently, he advised us to adopt a rule against lying and to enforce the rule. Each individual act could be more quickly assessed, not based on its own utility, but by its conformity to the rules.[4]

Applying Mill's defense in this case, the rule utilitarian can argue that we have reason to adopt a rule against punishing the innocent. People have reason to expect that others are trying to decide whether they are to be punished by deciding whether they are guilty of the crime, not that they are deciding punishment based on a utilitarian calculation of overall benefit versus harm (ignoring guilt or innocence). In fact, we would lose much of the deterrence effect of punishment (and the benefits that come from it) if people came to realize that their punishment was decided based, not on whether they committed the crime, but whether the punishment would produce a benefit. A rule that says to punish the guilty – and not to punish the innocent – provides the best way of harvesting these benefits.

Each individual act of punishment – such as the punishment of the police officer in the example above – is then to be evaluated according to whether it conforms to the rules, including the rule against punishing the innocent.

---

[4] Mill, John Stuart (1863), *Utilitarianism,* retrieved from https://www.utilitarianism.com/mill2.htm, 11/23/2016.

C. Boonin's Criticism of the Utilitarian Response

Boonin raises two problems with this rule-utilitarian response.

First, if it turns out that punishing the innocent is sometimes useful, then we should be able to come up with a rule that captures at least some of this usefulness. Instead of a broad rule against punishing the innocent, we could adopt a rule that prohibits punishing the innocent except in some specified set of circumstances that define when punishing the innocent is generally useful.

> *After all, if more utility is sometimes produced by punishing innocent people, then the rules that will produce the most utility apparently would have to be rules that sometimes permit the punishment of innocent people. (Boonin: p. 65)*

It seems quite likely that, since punishing the innocent sometimes produces more utility, that we can come up with a rule to capture at least some of the beneficial instances of punishing the innocent. (Boonin: p. 67) Even if it would be difficult in fact to come up with a set of rules that sometimes punish the innocent, it would still be the case that, if we were to discover such a set of rules, then it would be permissible to punish the innocent in those circumstances. The fate of our innocent police officer depends on it not being the case that a rule that allows punishment to prevent a riot produces more utility.

Second, Boonin argued that the rule utilitarian needs to explain why it would be wrong to violate the rules when it produces good effects. He calls this "rule worship problem" (Boonin: p. 69).

Boonin mentions a baseball player who is running from one base to another. A rule of baseball says that he must stay within the baselines. However, if he sees a child in the stands choking on her food, and he is the only person available who knows the Heimlich maneuver, he may break the rule and save the child. It is true that he is no longer to be counted as "scoring a run" in the context of the game. However, failure to "score a run" is of little consequence when compared to saving a child. (Boonin: pp. 70-71)

By analogy, a rule utilitarian needs an argument against stepping outside of the rule against punishing the innocent when she sees that it produces good consequences. It is true that the situation will no longer conform to the rules. However, if it produces better consequences, the fact that it violates the rules seems irrelevant.

The rule utilitarian, then, needs to explain why we cannot have rules that call for punishing the innocent when it produces an overall social benefit, or why we cannot break the rules against punishing the innocent when that produces an overall social benefit.

D. Application to Motive Utilitarianism

These objections challenge the rule utilitarian defense of punishment – and they are powerful objections against that theory. It seems, at first glance, that they would be equally telling against a motive-utilitarian theory of punishment.

First, if punishing the innocent sometimes produces good outcomes, then it seems plausible that we can come up with a set of motives that will call for punishing the innocent when it produces those good outcomes. In other words, we should be motivated to punish the innocent in those circumstances where being so motivated will realize the benefits of punishing the innocent. (Boonin: p. 77)

Second, the motive utilitarian needs to explain why, when the agent can produce the best possible outcome, it would still be wrong for him to do that rather than act according to what would otherwise have been the best motives. (Boonin: p. 78)

In the next section I will start to build a motive utilitarian theory of punishment with an eye to showing how it handles these and other objections.

III. Punishment in the Psychological Sense

I am going to leave the subject of criminal punishment for a while and talk about a different kind of punishment – punishment in the psychological sense. Punishment in the criminal sense, I will argue, is a special category of punishment in the psychological sense. I will use some important facts about punishment in the psychological sense in constructing a motive utilitarian theory of punishment in the criminal sense.

Punishment in the psychological sense is any state of affairs that a creature will expend energy to avoid. Bee stings, the upset stomach from a poisonous mushroom, electrical shocks, hunger, pain, loud noises, excessive heat or cold – all count as types of punishment. They are all ‘harms’ (in a suitably loose definition of the term), but they need not be intentionally inflicted. They do not even have to be inflicted by an intentional agent. However, harms intentionally inflicted on a criminal because he broke the law, does qualify as one type of punishment in the psychological sense.

If a rat pushes down on a lever, and that results in the rat suffering an electrical shock, then this shock I considered “positive punishment”. As a result of the electrical shock, the rat is expected to be less likely to press that lever in the future. One of the reasons for this is because the rat learns not to press the lever as a means to avoiding an electrical shock – standard deterrence. However, I would like to suggest that the rat also learns dislike pressing the lever. Avoiding the lever not only becomes a means of avoiding an electrical shock, it becomes an end in itself.

The description here matches claims that John Stuart Mill made with respect to certain goods such as virtue, money, power, and fame. These goods begin by being valuable as a means towards some other end – happiness. However, by means of their constant association with happiness become valued for their own sake. They become ends in themselves.[5]

In the case of virtue, Mill transitioned from discussing the utility of virtue to the utility of a love of virtue. These are not the same thing – any more than “my wife” and “my love for my wife” are not the same thing. Similarly, in our pet rat, we may distinguish between the rat’s aversion to getting an electrical shock, for the sake of which she avoids pressing the lever – and the rat’s aversion to pressing the lever, both of which are possible outcomes of the punishment of receiving the electrical shock.

An animal eats a particular type of mushroom and gets sick. It takes a lot of mental processing power to remember that this particular type of mushroom causes one to get sick, and to avoid being sick, one should avoid eating this type of mushroom. It would be more efficient to simply acquire an aversion to eating that type of mushroom. By its look, or its smell, or some other sensible property, the animal recognizes it simply as “that which is not to be eaten.” The animal that learns that aversion does not need to remember that the mushroom would make him sick. The animal that inherits the aversion (because ancestors that did not have it ended up dying) does not even have to experience getting sick.

Once animals acquire this capacity to acquire ends (such as an aversion to eating certain types of food) from experience (getting sick), then animals acquire something that other animals can use to alter their behavior. By arranging for an animal to be punished when that animal behaves in ways that are harmful to the agent, the agent can form in others aversions to performing those types of actions, and thereby avoid similar types of harms in the future. A bee’s sting, and

---

[5] Mill, John Stuart (1863), *Utilitarianism,* retrieved from https://www.utilitarianism.com/mill4.htm, 11/24/2016.

an ant's bite, are punitive in this sense. They are seldom fatal, but they do teach the animal a lesson.

The lessons of punishment go beyond those who are actually punished. If one animal observes another animal being punished, then the observer will acquire the same hesitation to perform the type of act that resulted in punishment as the one who was actually harmed.[6] So, the animal that punishes another animal to benefit from its changed behavior also benefits from the changed behavior of other animals of that type who either witnessed the punishment, or who witnesses the behavior that the punishment helped to bring about.

Among humans, punishment need not be actual or observed. Observational learning also takes place where the punishment is merely hypothetical – where nobody actually breaks the rules and, consequently, nobody is actually observed being punished. A culture that provides a threat of punishment for a particular type of act still generates an association between performing the act and the punishment in the minds of those within the community, and forms an aversion to performing that type of action.[7]

Furthermore, what counts as "observational learning" can be quite distant from the actual event. It can come from reading a news article or hearing a second-hand report. Albert Bandura and Richard Walters argues that humans are also able to engage in "observational learning" symbolically – by observing the behavior of characters in a work of fiction – the moral lessons that appear in a book or television show.[8]

A particular example of punishment in the psychological sense that researchers are interested in shows up in what is called *The Ultimatum Game*. In this game, one person is given a sum of money – say, $10. He needs to decide how much of this to give to a second person. He will have a chance to make an offer. If the other person accepts the offer, then he gives that amount

---

6 Mineka, S. & Cook, M. "Mechanisms involved in the observational conditioning of fear". *J. Exp. Psychol. Gen. 122*, 23–38 (1993).

[7] Lindström, Björn; Olsson," Mechanisms of social avoidance learning can explain the emergence of adaptive and arbitrary behavioral traditions in humans*." Andreas Journal of Experimental Psychology: General*, Vol 144(3), Jun 2015, 688-703

[8] Bandura, A., & Walters, R. H. (1977). *Social learning theory*.

of money to the other player and they both keep their amount. If, on the other hand, the other person refuses the offer, then nobody gets to keep anything. The original player must hand back all the money.[9]

When subjects are offered an opportunity to play this game, research shows that if the offer is less than about 30% of the total (e.g., the first player offers the second player 30% or less of the total), then the second player will often reject the offer. This is done to punish the first player for being unfair. The second player makes this choice even though accepting the offer would still have made him better off than he would have otherwise been, is done to harm (or, at least, deny a benefit) to the first player, and is done with a message of condemnation. It is also important to note that the player does this at a cost to herself. Making sense of these decisions under any consequentialist theory would be a challenge.

The decision to punish makes more sense if we look at it, not as an act of retributive punishment for this unfair offer, but – like the bee sting or the ant bite – as a way of creating an aversion to performing this type of action in the agent and in others not directly punished. The punisher might never see the other player again – and might not even know who it is if the game is conducted anonymously. However, that player will leave the game and go out into the community. There, not only will the player have learned a lesson, but others will learn through observational learning to be fairer with others in their own dealings. This situation must be compared to the alternative, where the player goes out into the community and brags about having gotten a large amount of money by offering the other so little (and he took it). This – through observational learning – teaches others to also be more selfish, with the resulting potential costs not only to the second player but those who the second player cares about.

Promoting an aversion to this type of action – the unfair distribution of such an award – goes far outside the context of the game.

These reasons to punish the first player are also reasons to demand that others who find themselves in a similar situation also punish the player. After all, if they accept a low offer, then they, too, will be creating a society with a weakened aversion to treating others unfairly. Furthermore, if one is going to condemn others for failure to punish, then this suggests that one would also have to condemn oneself. This leads to the attitude of, “I have too much self-respect to accept his offer,” and “that was not an offer – it was an insult.”

---

[9] Wikipedia, “Ultimatum Game”, https://en.wikipedia.org/wiki/Ultimatum_game, retrieved 11/24/2016.

Punishment in the Ultimatum Game still lacks two components of criminal punishment – one which Boonin mentions and one which he did not. The second player in the Ultimatum Game is not somebody with an authority to punish, which was one of Boonin's criteria for criminal punishment. In addition, the second player has no authority to escalate punishment to the point of violence. He may refuse the money, and nothing more. However, at this point, we are getting quite close to the types of punishment that Boonin argued could not be justified.

However, it gives us a model for understanding punishment in a motive utilitarian sense. The purpose of punishment – or the threat of punishment – is to create in the community at large an aversion to performing certain types of actions, such as theft, vandalism, assault, rape, and murder. Creating a community in which others have these aversions means living in a community where one does not have to worry as much about being the victim of theft, vandalism, assault, rape, and murder – and where one does not have to worry as much about one's friends and family suffering the effects of these types of actions. The tool for creating these aversions is punishment in the psychological sense.

The next step is to turn these claims about punishment in the psychological sense into a motive utilitarian defense of punishment in the sense criminal law sense. The next step in this process will be to apply what I have said so far to Boonin's criteria for criminal punishment.

IV. Punishment in Boonin's Sense

I have suggested that punishment in the psychological sense is about promoting aversions to certain types of behavior – more specifically, aversions to certain types of actions. Boonin was concerned with a more specific type of punishment – punishment in the criminal sense or punishing a person because he broke a law. Boonin provided us with a specific set of criteria for the type of punishment he had in mind. I would like to go through his criteria and describe how punishment in Boonin's sense relates to punishment in the psychological sense.

A. Harm

First, punishment involve harm or, as Boonin puts it, is "in some way bad for the person on whom it is inflicted." (Boonin: p. 5) This matches the theory of psychological punishment – where punishment is defined as subjecting a being to something that the being is willing to expend energy to avoid. Psychological literature further distinguishes between positive punishment – inflicting pain or some other unpleasantness on a creature, and negative punishment – depriving the agent of something the agent wants such as food, money, or liberty.

On this matter, there is no difference between punishment in the psychological sense and punishment in Boonin's sense.

B. Intention

Second, Boonin is interested in punishment intentionally inflicted. (Boonin: p. 11)

As I specified earlier, punishment in the psychological sense need not be intentionally inflicted. It does not even have to be inflicted by an intentional agent. It applies, for example, to feeling sick after eating a poisonous mushroom. However, Boonin is interested in the morality of punishment – and that means punishment that agents choose to inflict. Or, more precisely, it means "things that are in some way bad for another person" done intentionally.

C. Retribution

Boonin also reported that punishment contains an element of retribution. "[T]o be a punishment, an act must involve intentionally harming someone because he previously did a prohibited act." (Boonin: p. 16)

If a purpose of punishment is to promote an aversion to performing a certain type of action, then it makes sense to apply punishment to those who perform that type of action. To promote an aversion to lying, we punish those who lie. To promote an aversion to vandalism, we punish vandals. To promote an aversion to theft, we punish thieves. If punishment was not associated with what the agent did, or did not do, it is hard to imagine how it could be used to construct a social norm surrounding that type of behavior.

On this matter, I want to pull one item off this list and set it aside for separate consideration – punishing somebody "because he broke the law."

Boonin asserted repeatedly that the person who defended punishment had to defend the idea that a person was to be intentionally harmed "for breaking the law". However, the motive utilitarian theory suggests that this is not actually what is going on in many cases. The criminal is not being punished for breaking the law. She is being punished for theft, for vandalism, or for murder. To say that she is being punished for breaking the law would be to say that we want her to acquire an aversion to breaking the law – but not necessarily an aversion to theft, vandalism, or murder.

This is problematic because, if this is what is required, the person who defended a theory of punishment would have to justify punishing a person who broke an unjust law. It would have to

be a theory that said that punishing the person who hid Jews in her attic, or the person who helped runaway slaves escape from the south before the civil war. Boonin avoids these types of counter-examples by stating that a theory of punishment need only justify punishing a person "a just and reasonable law". However, this begs the question against the possibility that the reason that punishment is justified in some cases is not because the person broke the law, but for some other reason – the very same reason that makes the law calling for punishment just and reasonable.

The philosophy of law recognizes a distinction between acts that are *malum prohibitum* as opposed to *malum in se.* That is to say, actions that are wrong because they are illegal (no U turn) as opposed to acts that are wrong in themselves (no rape, no vandalism). We may find that distinction here – in cases where punishment aims to promote an aversion to the thing in itself (theft, vandalism, murder), and where punishment aims to promote an aversion to breaking the law or breaking the rules, such as traffic rules, which is accompanied by a list of rules not to be broken.

This does not impact Boonin's definition of punishment to this point. Whether we are talking about *malum prohibitum* or *malum in se*, we are still talking about intentionally harming a person because of something she did. It simply refuses to restrict the "something she did" to breaking the law.

D. Reprobation

Another criterion of punishment Boonin mentions is that punishment must contain an element of admonition or condemnation of the person punished. "[T]o count as a punishment for an offense, the act must express official disapproval of the offender." (Boonin: p. 21)

Insofar as we are seeking to create a general aversion to a type of act such as vandalism, it makes sense to inflict punishment along with the message that vandalism is something not to be done – that we disapprove of people performing that type of act. It is a true statement – and a statement that reports the reason why we are inflicting punishment. It works not only for cases in which we disapprove because a person committed an act of vandalism or theft – actions that are *malum in se,* but also when we disapprove of a person for the person for breaking the rules, thereby committing an act that is *malum prohibitum*. In the latter case, we object to the agent not having sufficient respect for the rules.

E. Authorization

Finally, Boonin argued that punishment only comes from a legitimate authority.

Of course, this is not true of punishment in the psychological sense. Inanimate nature can inflict punishment in this sense, and it still counts as punishment. Here, Boonin specifies that the topic of his conversation is “legal punishment” or punishment that is consequent on breaking the law.

However, this definition of punishment is going to lead to a problem. Boonin is going to want to argue that punishment in the sense that he is talking about – punishment that is “carried out by an authorized agent of the state acting in his or her official capacity,” (Boonin: p. 23) is impermissible. However, there are examples of punishment that contain all of the other elements of punishment that Boonin has mentioned, which seen clearly permissible.

Take, for example, the player of the Ultimatum Game who refuses an offer in order to punish the other player. The player, in this case, harms another person (by denying that person the money she could have otherwise kept), intentionally, because of something the other player did, accompanied by an expression of disapproval. The only difference is that the player is not an authorized agent of the state acting in her official capacity.

The question then arises, would it be wrong for an official of the state acting in her official capacity to engage in the type of punishment that somebody playing the Ultimatum Game seems morally permitted to do?

There are two ways of resolving this conflict. One is to say that being an official of the state puts one under special obligations and duties that prohibit the individual from doing that which others may freely do. As a private individual, I may show favoritism towards my friends and family on any number of matters, but not when I am acting as an official of the state. But I am at a loss as to see what would be wrong with empowering an agent of the state with refusing an offer in an Ultimatum Game when acting in her official capacity.

The other option would be to take Boonin’s challenge as applying as well to cases such as that of a person playing the Ultimatum Game – that the punishment he inflicts on the first player by refusing the offer is as morally suspect as the state’s punishment of a criminal. The challenge is to explain how either, or both, can be justified.

V. Another Type of Consequence

In discussing the good consequences of punishment, utilitarian theories have tended to focus on deterrence. Punishment increases the cost of crime for those who get caught, which is expected to result in less crime. However, motive utilitarianism focuses on another consequence of punishment – that of creating an aversion to performing certain types of actions. When we add these benefits onto the deterrence benefits, it raises the value of punishment, which contributes to its justification in utilitarian terms.

On standard deterrence theory, let us assume that an act is punished by a fine of $1000. Let us further assume that the agent holds that his chance of being caught is 5%. This means that the deterrence value of this law is $50 (or 5% of $1000). If the agent gets something from the crime that she would be willing to pay $60 to acquire then the fine provides inadequate deterrence. The rational agent would – and, in a sense, should – perform the illegal action.

Assume now that an agent acquires a moral aversion such that, “I wouldn’t do that even if you paid me $1000.” Let us keep the assumption that the agent believes that she has a 5% chance of getting caught. However, this assumption is irrelevant. The agent who is so averse to performing an act that she would not do so for $1000 is not going to give a different answer based on the odds of getting caught.

In cases where the chance of getting caught drops to zero, we lose the deterrence value of punishment entirely. However, moral aversions will continue to operate even under these conditions – providing the only motivational force left to prevent the agent from committing the crime.

The best protection that one can have against being murdered, raped, robbed, or assaulted is not a fear of punishment. We get much better protection if we can engineer our society so that people generally have a strong aversion to murdering, raping, robbing, or assault others. Legal deterrence – stopping people from committing a crime because they do not want to do the time or pay the fine – is, at best, a backup plan.

However, this does not help us to address the problem of punishing the innocent. The fact that we get more benefits from punishment than utilitarians usually discuss only means that we sometimes have more reason to punish the innocent – unless we can find a reason not to.

VI. An Aversion to Punishing the Innocent

It is now time to turn my attention to the claim that utilitarian theories justify punishing innocent people.

The motive utilitarian can handle this objection in a straight-forward manner – by introducing a motive against (an aversion to) punishing the innocent. The right act is the act that a person with the best motives to do. If those best motives include an aversion to punishing the innocent, then it is unlikely that there will be a case where punishing the innocent is the right thing to do.

Recall that Boonin raised the objection that, since punishing the innocent will sometimes produce the most utility, and since producing the most utility provides the ultimate justification for a utilitarian theory, then it seems likely that the best motives will, somewhere, be able to capture some of those good consequences. include justifications for punishing the innocent. consequences provide the ultimate justification in a utilitarian theory.

This appears quite likely in the case of rules. However, motives are significantly different from rules in several ways that have important implications for our ability to find a motive that justifies punishing the innocent.

A. Simple

Rules can be extremely complex. However, this is not the case with motives.

We can effortlessly create a rule with any number of exceptions and special circumstances built in. In chess, for example, a pawn can move only one square forward – and only directly forward - except on its first move, in which case it can move two squares, and except when it captures an opponent’s piece, which it can only do by moving on a forward diagonal, and if it should end its turn in the opponent’s home row the player may exchange it for a piece already captured.

Using punishment in the psychological sense to create a motive with this type of complexity would be difficult. It becomes significantly more difficult when we include the fact that we would be trying to create this complex motive across an entire population – as a universal moral sentiment. It may even be considered impossible.

One of the forms of "punishing the innocent" that Boonin raises against utilitarian theories are cases of vicarious punishment. He uses an example of punishing a child in order to deter a parent – somebody who is generally more interested in the welfare of their child than even in their own welfare. This may be a more effective form of punishment than punishing the parent

and, where the only thing that matters are the consequences, this may be the best form of punishment.

A motive utilitarian can respond to this by asking whether it is even possible to create a population that is indifferent to the suffering of an innocent child in this case without promoting a general indifference to the welfare of children. This is a psychological question – to be settled by empirical research. Yet, unless and until that research comes it, it is reasonable to worry that, in promoting vicarious punishment, we would tend to also create an indifference to the welfare of innocent people that we have to worry about the results.

The motive utilitarian will also have to ask about the effects of this form of punishment on the children. Having been harmed for the sake of controlling somebody else – or having their friends harmed for the sake of controlling somebody else – or hearing stories in which the heroes harm innocent people for the sake of controlling others, it seems reasonable to worry that the child will grow up to be somebody willing and even eager to harm innocent people as a way of controlling others.

At some point, the motivation will simply become too complex to be of a type that we can reliably teach to a whole population. At some point, the question passes from, "Is this practical?" to "Is this even possible?"

These questions do not apply to rules – which can be made as complex as we wish. The teaching of a rule involves nothing more than telling it to them. They provide limits to what is practical or even possible within a motive utilitarian framework that a rule utilitarian theory does not confront.

One could raise the objection, "But, you are still saying that if it is possible to create such a complex aversion, then it may be permissible to punish the innocent. Since this conclusion itself is unacceptable, it is enough to call for rejecting the motive utilitarian theory."

It does imply this. However, to evaluate the claim that this conclusion is unacceptable we should look at an example where we do have an exception.

The aversions that we teach includes an aversion to taking the property of others without their consent. However, we build in an exception for taxation. Setting aside political theories that say that we consent to a social contract that includes taxation, taxation generally involves somebody taking one's property under a threat of violence.

One of the effects of building in this exception is that we generate confusion. There are those who take this aversion to taking the property of others without their consent and assert that taxation is theft. There are others who say that the state or the people can take property for the public good without limit. A simple aversion to taking property without consent would avoid these confusions. However, the benefits of taxation override the problems with including an exception.

More to the point, if there were a case in which we should have no aversion to punishing the innocent, then we would likely regard it as we do taxation. I suspect we would not call it punishment – reserving that word for the case in which we punish the wrongdoer (preserving the simple claim that it is wrong to punish the innocent).

However, we could build it into our institutions under a different name; in the same way that we use different names for "theft" and "taxation", and different names for "slavery" and "conscription". In each of these pairs of cases, the latter may be understood to be a useful application of the former, wrapped in rules that increase the overall benefit while decreasing the overall cost

Utilitarianism is said to justify slavery in some circumstances. However, that slavery would have to be a type of slavery wrapped in practices and rules that would produce the best consequences. If, in our slave culture, we could improve overall utility by providing our slaves with some small amount of money that they can spend for themselves, a utilitarian would require that slaves be paid a minor compensation. And if it should be the case that people make the best slaves while in the prime of their lives, perhaps our slave culture would be improved by enslaving each person for a few years in the prime of their lives, then by enslaving a few people for their whole lives. When we are done wrapping slavery in all of the rule, practices, and procedures that would maximize its benefit and minimize its costs, we may then give it a new word (conscription; compulsory community service) to distinguish it from forms of slavery – properly so called - that lack utilitarian refinements or justification.

Consequently, when we are presented with an intuitive objection that utilitarianism would justify slavery, we are then invited to imagine slavery without these utilitarian refinements – which utilitarianism does not justify. At the same time, we are lead to ignore the utilitarian improvements we could make to what we imagine. When slavery comes wrapped in these utilitarian refinements and is given a new name - “conscription” or “compulsory community service” – we find that it is not so intuitively objectionable.

In fact, there is a type of "punishment of the innocent" that we do tend to accept – and we even call it “punishment”. These are cases when we punish a group of people for the infractions of some of its members. A teacher may cancel an event for the whole class because some of the students in the class are misbehaving – thus punishing the whole class, including those students who are innocent. Similarly, a military officer may require a unit to undergo additional drills because some of its members committed an infraction. Given these examples, I think there is reason to doubt that, if we found other examples in which punishing the innocent produces a benefit, and we wrap it in rules and procedures that minimize the cost and maximize the benefit, that we would call it wrong.

B. Persistent

We may create a rule that says, “only eat fish on Friday.” However, we cannot (or, at least, cannot without a great deal of effort and a low probability of success) create a motive that says, “only like fish on Friday”.

Once an aversion is created, it will remain in force through all of the circumstances where it is applicable. There will be no “turning it off” in a specific case simply because some utilitarian good can come from it.

Consider a person with a fear of flying who must travel across the country quickly to deal with an emergency. The fear of flying is still going to be there, motivating her to find some other way to travel, or to fly without experiencing it (knocked out by drugs), or to avoid flying and handle the emergency – even if less efficiently - from her current location. If the aversion is strong enough, she might say, “I can't do this. You are going to have to find someone else.” The option of turning the aversion off simply because, in the current circumstance, it is getting in the way, does not exist. If the fear of flying is strong enough, an individual might simply have to say, “Look, I can’t do this.”

Consider next a person with an aversion to punishing the innocent. Even if something comes up where there is some benefit to punishing an innocent person, she will not be able to simply shut off the aversion. It will push her to search for some other option. If no other option is available, she may still judge that the other reasons are not good enough. Depending on the strength of the aversion, she may even say, “I just cannot do this.” It will take overpowering reasons to get her to punish the innocent.

I offer it as a point in favor of this account that, in spite of the prohibition against punishing the innocent, if the stakes were high enough, we would expect a person to go ahead and punish the innocent – though we would also expect him to feel terrible about doing so.

For example, if terrorists are threatening to set off nuclear weapons in several major cities unless the government were to frame an individual for a crime and have that person imprisoned or executed, this may be the right thing to do. The motives to save life and reduce suffering may be strong enough with so many lives and so much suffering at stake to override the aversion to punishing the innocent in this circumstance. However, the aversion to punishing the innocent – like the fear of flying – will not simply disappear. It will remain, causing the agent to dislike what he is being forced to do – a dislike that will motivate him to struggle to find some way to get out of it, having a residual effect on the emotions of the agent forced to do something she strongly disapproves of doing.

In all other cases – including many cases when there are benefits to be had if one did not have a fear of flying or an aversion to punishing the innocent – the aversion will be enough to motivate the agent against taking the more useful option.

The one difference that exists between the fear of flying and the aversion to punishing the innocent is that the former is an aversion that people generally have little or no good reason to promote, while the latter is an aversion people generally want those who surround him to have. It is a reason he has reason to promote. To help promote this aversion, his best tools include cultural institutions that condemn and punish those who are caught punishing the innocent.

C. Motivational Force

Each motive comes with its own motivational force; rules do not. When we give a person a rule, we still have to answer the question, "Why should I follow this rule?" However, when we give somebody a motive, the answer to the "why" question is built into the motive itself.

There is a different question to answer with respect to motives that is, "Why should I have this motive?" We have answered this question in terms of the reasons that exist for using punishment – in the specific set of cases we are currently concerned with – to create in people certain aversions. However, once the motivation exists, it will necessarily push the agent to perform certain actions and refrain from others according to the strength of the motive.

The claim that motives (desires and aversions) come with built-in reasons relates to Bernard Williams' account of what it is to have a reason.

A has a reason to φ iff A has some desire the satisfaction of which will be served by his φ-ing[10]

An individual with an aversion to punishing the innocent has a reason not to punish the innocent – a reason whose strength is determined by the strength of the aversion. If an agent has a particularly strong aversion to punishing the innocent, then it will take a great deal of weight to cause him to set it aside. No trivial increase in overall social utility will suffice.

D. Objects of Aversion

A person with an aversion to punishing the innocent will be motivated to avoid punishing the innocent for its own sake – and not for the sake of some other end, in the same way that a person with an aversion to pain will be motivated to avoid pain for its own sake, and not for the sake of some other end.

In other words, if a person has an aversion to punishing the innocent, then the reason she will give for not punishing a person is “because I would be punishing the innocent – and I will not do such a thing”. The reason she has an aversion to punishing the innocent may be justified by utilitarian considerations – because people generally have many and strong reasons to cause the members of society to have an aversion to punishing the innocent. However, once that aversion is built into an agent's motivational set, then "because I would be punishing the innocent – and I will not do such a thing" becomes the agent's reason for action.

Similarly, a person with an aversion to lying (brought about by punishing and condemning those who lie) will refuse to lie "because it is lying" - and a person with an aversion to vandalism will refuse to vandalize the property of others "because it is vandalizing the property of others".

When this person is asked, "And why is it wrong to lie?" or "Why is it wrong to vandalize the property of others?" she may have a hard time answering the question, unless it occurs to her to answer, "What reasons do we have to condemn and punish those who lie?" and "What reasons exist to condemn and punish those who vandalize the property of others?" The answer to these questions have to do with the many and strong reasons that exist for creating an aversion to lying, and creating an aversion to vandalizing the property of others.

---

10 Williams, B., 1979. “Internal and External Reasons,” reprinted in Moral Luck, Cambridge: Cambridge University Press, 1981, 101–13).

The same applies to the aversion to punishing the innocent. The person with such an aversion will refuse to punish the innocent because punishing the innocent is something he has an aversion to doing. When answering the question of why it is wrong to punish the innocent, that question comes from answering the question, "Why promote an aversion to punishing the innocent by condemning and punishing those who would do so?"

E. Acts and Omissions

The object of an aversion to punishing the innocent – just as with aversions to theft, vandalism, and the like – is an aversion to performing the act. It is not an aversion to the act being performed.

Consider what it would take to create a universal aversion to it being the case that there are innocent people being punished. Creating aversion to any situation in which an innocent person is punished would mean being punished whenever it is the case that an innocent person is punished – which would put us permanently in a state of punishment. There will always be cases in which innocent people are being punished – both in terms of the misapplication of good law and because some people who ought not to be punished live under bad law.

On the other hand, we create and sustain an aversion to punishing the innocent, as we do with an aversion to lying, breaking promises, vandalism, theft, and the like, by threatening to punish those who perform an act of lying, breaking promises, vandalism, theft, and of punishing the innocent. This is not nearly as problematic.

One of the implications of this comes out when an agent is faced with a choice where to perform an action or to prevent some similar fate from befalling a larger number of people. Utilitarian critics will bring up hypothetical cases such as that of killing one healthy patient to spread his organs out among five patients who would otherwise die of natural causes. We have reason to create in people a moral aversion to killing by condemning and punishing those who kill. However, to create a similar aversion to people dying of other causes we would have to condemn and punish people whenever others die of other causes. Even if we limit the latter to deaths that we have the ability to prevent, that list is endless.

F. Omnipresence of Motives

So far, I have described moral aversions acquired through punishment as simple, persistent rules against performing specific types of action (theft, vandalism, punishing the innocent) that come with their own motivational force.

It is also the case that, while we can step outside of a set of rules, we cannot step outside a set of motives.

In raising his objections to the rule-utilitarian defense of punishment, Boonin provides an example of a base runner in a game of baseball (Boonin: pp. 71-74). The rules of baseball state that in order for his act to count as scoring a run, he must stay within the baselines. However, if he sees a child in the stands who is choking on some food, and he alone knows the Heimlich maneuver, he might decide to go and save the child. When he does so, he steps outside the rules of baseball – which does not have a rule that covers the case of a child in the stands, choking. However, the runner must still have a motive.

The ball player's reason might have been to save the child. Or it might have been a public-relations stunt so that the fans will see him as a hero and increase his popularity (and the money he can draw from product endorsements). Unless he tripped and accidentally saved the child, he acted for a reason, even when he did not act in accordance with some rule.

So now, in this motive-utilitarian framework we have to ask, "What is this motive? What is its value?"

Saving a child's life is a good motive – one we have reason to encourage and promote by rewarding and praising those who act on such a motive.

Promoting one's popularity so as to make more money from product endorsements seems to involve motivation that people generally have no reason to either promote or discourage. However, the lack of an aversion to lying – exhibited by making up a fictitious public relations stunt with a child-actor pretending to choke – would be something that people generally have reason to condemn.

Similarly, a person with an aversion to punishing the innocent cannot simply leave it behind and step into a realm where motives do not apply. This is different from saying that the aversion to punishing the innocent persists. It is saying that some other motive must be brought into play to outweigh the motivation that this aversion provides. Furthermore, the desire to win a game of baseball is quite trivial, whereas the aversion to punishing the innocent is an aversion that we would want people to have in great strength. It will take more than a good reason to provide praiseworthy motive to override the aversion to punish the innocent. It would have to be an extraordinary reason.

This fact tells against Boonin's claim that following motives counts as a motive utilitarian version of "rule worship". The agent is no more "worshipping" motives than he is "worshipping" the laws of physics by choosing not to float away from the ground.

G. Public Teaching

Many of the arguments against act- and rule-utilitarian defenses of punishment bring forth examples where things are done secretly so as to avoid some of the harsh negative consequences of the act or of acting in accordance with the rule. For example, an innocent person is framed for committing a crime so that he may be punished, where that punishment will provide deterrence and produce good benefits.

However, when we talk about engineering a sentiment, those sentiments are to be engineered across the whole of society.

It would be difficult to train a whole population to have an aversion to "punishing the innocent except when one is a legislator considering a proposal to establish a secret tribunal or except when one is placed on such a tribunal". The problem rests not only with the complexity of such an aversion, but with the fact that people are to pick up this aversion in their enculturation- from the attitudes, expressions of approval and disapproval, and observed rewards and punishments (imagined and real) depicted around them. Rules can be adopted in secret and acted on in secret. Moral aversions cannot.

Furthermore, these moral aversions need to be taught to a diverse population with a range of experience and capacity – to the whole population except a few considered incapable of it. Morality is not meant for a small handful of intellectual elites, and an aversion that only a handful of elites can learn would not serve the needs of morality.

Once people acquire this moral aversion to punishing the innocent, this is going to found in the legislator considering a law to set up a secret tribunal, as well as to any person who sits on that secret tribunal. We would have to somehow teach people generally not to have an aversion to punishing the innocent when setting up or sitting on a secret tribunal for punishing the innocent. That would not be easily done. It would motivate them to avoid punishing the innocent because they are innocent, not to set up an institution to do that which they oppose.

H. Summary

Boonin actually only addresses the motive-utilitarian option with a single rhetorical question.

> *How could the move from act-or rule-utilitarianism to this motive-based version produce a more successful solution to the problem of punishment? (Boonin, p. 71)*

I hope the preceding section answers this question. Motives are different from rules in a number of important respects.

Rules can be extremely complex; motives must be simple.

Rules can be temporary or intermittent; motives are persistent.

Rules do not come with their own motivational force but must find the reason to follow the rule from something external; a moral aversion comes with its own reason for the agent to act or refrain from acting.

Rules do not create ends, but are always a means for achieving other ends; motives are propositional attitudes the objects of which, in virtue of the motive, become ends in themselves.

We can have rules about anything with equal ease; we can teach people to have aversions to performing certain types of action far more easily than we can for creating aversions to certain states of affairs – particularly states of affairs we cannot change.

Rules can be made in secret for secret reasons; moral aversions are taught to the public at large using the tools of enculturation.

Finally, rules come into existence with but a thought, while punishment and condemnation consistently and universally applied provides the tools for creating moral aversions.

With these limitations in place, it is not so clear that we can come up with something better than a strong and persistent aversion to punishing the innocent because they are innocent taught widely and publicly by condemning and threatening to intentionally harm anybody who punishes the innocent.

VII. Implications of an Aversion to Punishing the Innocent

Once we have this aversion to punishing the innocent, what do we do with it? What concerns will it motivate?

If an individual has an aversion to punishing the innocent, then that person is going to motivated to make sure that an individual is guilty before he is punished. This aversion will not so likely be strongly triggered if the harm is a small harm – like being yelled at or or inflicting some other

sign of disapproval. However, when the harms are more severe, and when they are harms that the state can escalate into.increasing levels of violence up to and including death, those not wanting to punish the innocent will see stronger reasons for safeguards.

The strategies that we can use to help to reduce the chance of an innocent person being punished may include a requirement that the evidence of guilt be presented to an impartial jury, and that the jury operate on a presumption of innocence. That is to say, we are going to create an institution that presumes that accused is innocent and ought not to be punished unless proof can be provided beyond a reasonable doubt. These, too, are a part of the aversion to punishing the innocent.

In fact, support for these safeguards becomes a test of one's aversion to punishing the innocent. The individual who cares nothing about a trial, nothing about evidence, lacks an aversion to punishing the innocent and is condemned for it. This condemnation becomes one of the ways that a culture promotes an aversion to punishing the innocent – by condemning or punishing those who do not act as a person with an aversion to punishing the innocent would act.

VIII. Additional Objections

Boonin raised other objections to motive utilitarian theories – and to other theories which, at first, may seem applicable to motive-utilitarian theories. I would like to address some of these objections Next.

I. Boonin's Requirements for Motive Utilitarianism

When Boonin brought up the theory of motive utilitarianism, he said that the motive utilitarian would have to do two things.

> *A proponent of a motive-utilitarian solution would have to establish two things: that having the motive to ensure that people are punished for breaking the law would produce more overall utility than would having any alternative motive, and that if this motive is the best one to have, then acting from this motive would always be the right thing to do. (Boonin: p. 78)*

I do not see this as being the requirement at all.

In writing about "the motive" and "this motive," Boonin seems to be constraining the motive utilitarian to the thesis that there can be only one motive for punishing the guilty. Perhaps this

motive is a desire to intentionally harm those who are guilty. Whatever this motive is, the motive utilitarian must show that this one motive is the best that an agent can have and that it is always permissible to act on it.

Yet, no agent acts solely on one motive. And what motivates an agent to do something is seldom a motive to do that thing. Instead, she is motivated to obtain some other end and the act is simply a means toward that end.

What the motive utilitarian needs to do is show that a person with good motives would have reason to punish the guilty.

The reasons to punish are easy to come by. They are the reasons people have to form a community where others are disposed to tell the truth, keep promises, and repay debts. It comes from the reasons they have not to have their property destroyed, not to be assaulted, raped, or murdered, and not to have their property taken. This is combined with the fact (if it is a fact) that a culture of punishing those who perform these types of actions create moral aversions to lying, breaking promises, theft, vandalism, assault, rape, and murder. This yields reasons to establish a culture that punishes these types of actions.

The motive utilitarian does need to concern himself with whether there are better reasons not to punish than to punish.

For example, an aversion to harming others would provide a motive against punishing others, since to punish them is to harm them.

Against this, we content with the fact that the simple fact that we are living and in competition with others in some areas means that we are going to harm them. A person who opens a business will harm its competitors. An individual who applies for and accepts a job offer leaves at least one other applicant worse off than she would have otherwise been. Recently, Uber tested a self-driving truck by having it make a beer delivery – an invention that will have the effect of harming a great many truck drivers who may find themselves out of a job. When a company fires an employee or closes a factory, workers are harmed, as well as businesses that sold goods and services to those workers. We may have reason to promote some aversion to harming others, but the necessities of life suggest it is an aversion that can be easily overridden by other concerns.

Even Boonin's own suggestion of replacing punishment with compulsory restitution involves harming the person compelled to make restitution.

However, these harms are side effects of performing other action. Perhaps what we need is an aversion to intentionally harming other people. Since punishment involves intentionally harming others, such an aversion would provide people with a reason not to punish.

As I have described above, the practice of intentionally harming others – at least harming them in the psychological sense – goes far beyond state punishment. It applies to the player in The Ultimatum Game who punishes the person who makes too low of an offer by refusing it. It applies to decision to boycott a business that has adopted a rule or policy that potential customers find objectionable. It applies to the regular criticisms and condemnations of daily life. The tone that a driver uses when shouting at a pedestrian for jaywalking, or a parent uses when condemning a child for writing on the walls are used, if this account is correct, for the same reasons that criminal punishment is used. The unpleasantness of the situation acts on the reward centers of the brain to form an aversion to the type of act that one condemns.

This brings up another problem with an aversion to intentionally harming others; how do we create such an aversion? If this account is correct, we intentionally create and strengthen moral aversions in others by inflicting what are punishments in the psychological sense – associating the type of action with a state that individuals will typically expend energy to avoid. Attempting to create a moral aversion to intentionally harming others would seem to require that we do something that the very state we are trying to create would prohibit us from doing. Perhaps a day will come when we can create these aversions without using punishment in the psychological sense. When that day comes, we may have reason to give everybody an aversion to intentionally harming others. But that is not the situation that we find ourselves in today.

The motive utilitarianism has reason to reject the aversion to harming others (which we have to do just in the course of living), and the aversion to intentionally harming others (which would eliminate our ability to create moral aversions at all, including the aversion to intentionally harming others). The option that works, in this regard, is an aversion to intentionally harming the innocent. This is an aversion that we have many and strong reasons to promote universally and publicly by punishing (in the psychological sense) those who punish the innocent. But this not only fails to provide a reason not to punish – it is a reason to punish those who perform certain types of actions – to build and strengthen that aversion.

J. Not Punishing the Guilty

Another set of objections that Boonin raised against several of the theories that he considered is that the theory calls for the punishment of people who we do not, intuitively, see as deserving of

punishment. For example, in the case of the moral education theory of punishment, Boonin pointed out that there may be good evidence that a person needs a moral education, but has not necessarily performed a criminal act. Cases like these raise problems with the idea that moral education justifies punishment.

Against motive utilitarian theory, one can try to make a similar case.

Consider, for example, a person who breaks a promise to meet a parent for lunch. We have many and strong reasons to promote an aversion to breaking promises. On the motive utilitarian theory presented here, we have reason to punish those who break promises as a way of creating those aversions. Yet, in many cases, breaking promises lies well outside of the law. If one person breaks a promise to meet another person for lunch, this hardly seems like a matter for the police.

Lying and promise-breaking often are not subject to legal punishment. There are a few exceptions – such as lying under oath – but these exceptions do not change the fact that an individual can seldom call in the police when a roommate forgets to buy milk as promised, or a husband tells a wife that he was working late.

However, it is important to remember that motive utilitarianism, as a utilitarian theory, is concerned with practical matters. The motive utilitarian can address the objection by pointing to the consequences of having these simple lies and breaking of promises written into the criminal law. If the state were to get involved in investigating every instance of lying and promise-breaking that takes place – every employee who is late for work, every friend who fails to show up to help move – it would require a massive increase in the size of the police force and the court system. The costs would be prohibitive. The benefits, even though potentially real, are not large enough to justify these costs.

Prudence dictates that the punishment for the act-types of lying and promise-breaking remain substantially in the realm of private action. This would limit the types of punishment permitted to those that individuals can carry out on their own – verbal condemnations and refusing to interact with the liar or promise-breaker. It prohibits that feature of punishment that is limited to the state – the authority to escalate punishment up to the level of violence and even death.

Another type of case where we make certain actions immune to legal punishment, even when there are reasons to condemn or punish the actions, is in the realm of freedom of speech. In expressing certain ideas, we can sometimes reliably determine that the speaker has attitudes

that people generally have many and strong reasons to condemn. Motive utilitarianism, it seems, would call for the punishment of these people. Yet, we hold that it is inappropriate to punish them.

However, John Stuart Mill in *On Liberty*[11] provided us with many consequentialist arguments against including speech in the realm of actions that we should respond to with violence. Primarily, this is because cultures that have punished people for words alone have tended to punish people for expressing ideas that were actually true and beneficial to society. The risk of punishing people for expressing beneficial ideas has been judged to exceed the benefits of not punishing people who express bad ideas, so we have (at least in some parts of the world) prohibited the practice of responding to mere words and communicative actions with violence.

Mill also argues that certain actions – lifestyle choices – that concern only the individual and that violate the rights of nobody else, though they may be offended by or worried for the agent, also should not be met with the type of punishment that can escalate to violence, but be left to the private admonitions of concerned individuals. Mill argues that making them a concern of the state will do more harm than good.

A third gap between who motive utilitarianism identifies as deserving of punishment and yet our moral intuitions judge as not to be punished are those who performed an act type that people have reason to condemn, but which has not (yet) been made criminal. The motive utilitarianism needs to justify the fact that we cannot actually punish a person until after their act has been made criminal, and only then if they repeat the act after the law has gone into effect.

The arguments here are the arguments against creating *ex-post facto* laws. History shows that where political leaders have the power to make act-types criminal after the fact, they have tended to use this power to declare the past act of a political rival to be criminal and use that to have the rival arrested and punished. A prohibition on *ex-post facto* laws exists to prevent these types of abuses. One of the costs of this type of restriction is that a person can commit an act type that people generally have many and strong reasons to promote an aversion to performing, yet escape punishment up to the point that the community actually makes the act-type illegal.

These examples show how, in general, the motive utilitarian would respond to the accusation that the theory calls for the punishment of people that we seem to think ought not to be

---

[11] Mill, John Stuart, *On Liberty*, retrieved from: http://www.bartleby.com/130/, retrieved, 12/01/2016.

punished. It looks at the consequences of using an alternative set of principles and judges if there are any that would actually do a better job.

K. Punishing the Innocent

There is a flip side to the issue of a theory calling for the punishment of people that we intuitively see as not deserving to be punished. This is for the theory to deny that punishment is legitimate where we intuitively see punishment as being warranted.

Boonin repeatedly used an example of a person who drives a sick friend to the hospital in the only vehicle available – one that has failed to pass its emissions tests. This person does nothing wrong in a motive-utilitarian sense. In specific, we do not have good reason to cause people to form an aversion to offering such assistance to a sick friend – even to the point of violating a pollution ordinance. Yet, it is still the case that the driver broke the law and, as such, is liable for criminal punishment. The motive utilitarian cannot account for the criminal punishment of this law-breaker.

In this type of case, I am going to argue against Boonin's assumption that the motive utilitarian needs to explain why punishment in this case would be justified. I would argue, instead, that it is not justified. The agent in this case ought to be let off – perhaps with a warning, but not punished. The law should recognize that she acted for good reason – not the type of reason that deserves punishment – did not perform the type of action that we have reason for people generally to be averse to performing – and, thus, refuse to punish her.

There are several ways to create a legal system that respects the possible good reasons for performing an act-type that, generally, people have reason to give people an aversion to performing.

One option is to write the exceptions into the law itself. A law against speeding can include an exception for cases where there is a medical emergency that is a matter of life and death. Prohibitions against assaulting and even killing another person often come with a built-in exception in the case of self-defense or the defense of another.

Alternatively, the law can be written in vague terms that require those who enforce the law to apply moral judgments. The act being prohibited might be defined as "showing reckless disregard" or "with the intent to cause harm" that would exclude actions performed without such reckless disregard or with the intent to help somebody in need.

Yet another option is to give people discretion as to when to enforce the law. The victim's discretion to press charges, the police officer's discretion as to give a warning or citation or to ignore a criminal act entirely, the prosecutor's discretion to file charges, the grand jury's discretion as to whether to indict, the jury's discretion as to whether to convict, the judge's discretion as to overrule a jury's guilty verdict, and a governor's or president's discretion to pardon, are all examples where the law provides opportunity to prevent a good person from being punished. The moral aversion thesis suggests that we use these features to handle the cases that Boonin brings up whereby good people might find themselves being punished for performing that action that a person with good desires would perform.

There is an argument to be had against using this type of discretion, but it is not the type of argument that threatens the motive utilitarian theory of punishment. To the degree that we give people discretion in how they enforce the law, to that degree we risk having them act on explicit and implicit biases – allowing them to show favoritism to their friends and family, their business associates and neighbors, to their fellow church members, to their race or gender, or to active members of their political party. To prevent these types of abuses, one may argue for less discretion in applying the law rather than more.

In the same way that practical considerations say that we may be better off if we do not apply those types of sanctions that can escalate into violence against all liars, promise breakers, or speakers, it may say that we need to apply the law even to those who violate the law for good reason. The utilitarian costs of not doing so – in terms of favoritism and possible corruption – are too high.

Another type of case that tests our intuitions regarding the relationship between morality and law are areas where the law draws a bright line in a field of moral gray. The age of consent and the amount of money stolen that distinguishes a misdemeanor from a felony provide just two examples of cases where the law creates a bright line, where morality sees only slight variation in shades of gray. Consequently, where morality does not support the conclusion that the person who stole $500,00 deserves significantly more punishment than the person who stole $499.99, the law simply cannot be as fine-grained as morality.

In short, the motive utilitarian theory of legal punishment does not, in fact, show that legal punishment conforms precisely to the moral justification for punishment. However, this is because humans are fallible beings, the law is an imprecise tool, and people have a lot of things

to worry other than – and sometimes in competition with – using punishment to promote aversions to certain types of actions.

IX. Utility

I have argued for a defense of punishment based on a motive utilitarianism. However, the argument that I have provided has an important implication that threatens the utilitarian component of this argument.

With motive utilitarianism, motives are evaluated according to the degree to which they tend to increase total utility. Utility is typically understood in terms of pleasure over pain, happiness over unhappiness, satisfaction over dissatisfaction, or some similar good. On all cases, utility, however it is understood, is the sole good used in making these evaluations.

In the course of making this argument, I have claimed that one of the consequence of punishment is to create new ends. A culture of punishment regarding types of action such as lying, breaking promises, theft, vandalism, assault, rape, and murder creates aversions to these types of actions. These aversions create new ends. The individual with an aversion to lying is one for whom "I am not lying" becomes an end in itself. These become things that the agent seeks to avoid independent of the effects that telling the truth may have as a means to maximizing utility.

In the arguments given above, this was particularly relevant to the aversion to punishing the innocent, which is made an end in itself. As a consequence, the agent will not punish the innocent even when the innocent maximizes utility.

As agents acquire these new ends, or the strength of these ends are altered by the application of rewards (such as praise) and punishments in the psychological sense (such as condemnation), these ends become new reasons for action. As such, they become new reasons to reward or praise others, and new reasons to condemn or punish others. It is these ends, not utility, that provide the measure of (other) motives.

John Stuart Mill attempted to reconcile these consequences with utilitarianism by arguing that the new end becomes "a part of happiness". Speaking of the love of virtue, he wrote:

> *Whatever may be the opinion of utilitarian moralists as to the original conditions by which virtue is made virtue; however they may believe (as they do) that actions and dispositions are only virtuous because they promote*

> *another end than virtue; yet this being granted, and it having been decided, from considerations of this description, what is virtuous, they not only place virtue at the very head of the things which are good as means to the ultimate end, but they also recognise as a psychological fact the possibility of its being, to the individual, a good in itself, without looking to any end beyond it; and hold, that the mind is not in a right state, not in a state conformable to Utility, not in the state most conducive to the general happiness, unless it does love virtue in this manner- as a thing desirable in itself, even although, in the individual instance, it should not produce those other desirable consequences which it tends to produce, and on account of which it is held to be virtue.*[12]

It would make more sense to simply abandon this convoluted form of reasoning and simply admit that, when reward and punishment creates new motives, it creates new ends, distinct from utility, and sometimes in conflict with it. This ultimately leads to an abandonment of utilitarianism strictly speaking. However, what is left – a pluralistic form of consequentialism where every desire and aversion creates its own end - is not significantly different from what Mill was trying to defend.

At first glance, this may be thought to create a circular form of reasoning as motives are evaluated according to other motives. However, it would be more accurate to view this not as a circular argument, but as a feedback loop. A set of desires provide reasons to reward and praise some actions and to punish and condemn others. Reward and praise, in turn, bring about changes in motivational states, which provide new reasons to reward or punish others.

By analogy, one can think of the way in which increasing global temperatures result in more evaporation of water. This evaporation means more water vapor in the atmosphere. Water vapor is a greenhouse gas, which traps in more solar energy, which in turn leads to still more warming. At the same time, this water vapor may form more clouds, which reflect sunlight, which would bring about a reduction in overall temperatures.

What we end up with, then, is not a theory that seeks to maximize utility, but a theory that seeks harmony among a plurality of ends. Each motive is evaluated by its ability to fulfill or thwart other motives, which provides the reasons and the motivation to reward/praise or punish/condemn.

---

[12] Mill, John Stuart (1863), *Utilitarianism,* retrieved from https://www.utilitarianism.com/mill4.htm, 11/24/2016

X. Conclusion

Why is punishment or condemnation the appropriate reaction to wrongdoing?

A great deal of philosophical work has gone into trying to discover what it is that justifies calling an action wrong – to come up with the distinguishing characteristics that distinguish what is wrong from what is permissible and what is obligatory. However, one question that seems to have been substantially overlooked is the question of, "Why is condemnation or punishment the appropriate response to wrongness?"

I argue that the reason for condemnation and punishment is because of the effects they have on behavior. These responses not only deter those who wish to avoid condemnation and punishment, but these actions act on the brain to form aversions to the very types of acts likely to result in condemnation and punishment. In acquiring an aversion to these types of acts, people tend to avoid doing them for their own sake, even when they may serve some other end. Being surrounded by people with an aversion to lying, breaking promises, theft, vandalism, assault, rape, and murder allows each of us to live a better life.

The justification for punishment comes from the justification for creating – across the whole population – these widespread aversions.

Boonin's objection that a consequentialist defense of punishment suffer from the implication that they would justify punishing the innocent is answered by pointing out the good consequences of an aversion to punishing the innocent. Moral aversions, unlike rules, must be simple, persistent, self-motivating rules that take their object (not punishing the innocent) as an end in themselves that are taught universally across a whole population. While it is almost certainly the case that we can come up with a rule that would call for punishing the innocent, it is much more difficult to argue against a universal aversion to punishing the innocent.

One of the implications of this claim that condemnation and punishment create aversions that make their objects ends in themselves is that we are forced to abandon utilitarianism as the ultimate justification for motives, rules, or actions. There is no single end (e.g., happiness) to appeal to in evaluating motives. There is a long list of ends, each the object of their own desire and aversion, to appeal to in making these evaluations. Motives that tend to fulfill other desires are those that people have reason to promote using praise and rewards, while aversions that prevent people from acting in ways that thwart other desires are those that people generally have reason to promote using condemnation and punishment.

A day may come when humans have some other way to create aversions to act-types such as lying, breaking promises, theft, vandalism, assault, rape, murder, and punishing the innocent. Until that day comes, the one tool that we do have is punishment. Giving up this tool would likely prove to be very costly.