The Real Problem with Prepunishment

Abstract

When someone is postpunished, they are punished for a crime they have committed. When someone is prepunished, they are punished for a crime they will or would commit. This paper targets two assumptions in the philosophical literature: 1) that prepunishment exists only in science fiction, and 2) that prepunishment is unproblematic for consequentialists. First, I argue that the differences between the actual world and hypothetical cases of prepunishment — even those presented in Philip K. Dick’s “Minority Report” — are smoke and mirrors. In fact, there is no morally relevant difference between “Minority Report” and how we often punish people for attempt offenses. Therefore, if prepunishment is morally wrong, as nonconsequentialist moral philosophers have argued, then we need to reform legal theory. Second, I argue that prepunishment is problematic for consequentialist reasons. Most importantly, prepunishment mechanisms have no deterrent power. The power that prepunishment mechanisms do wield, which I label “character control,” can, under certain conditions, lead to extraordinarily bad societal consequences. To avoid these negative consequences, we must ensure that our prepunishment mechanisms do not become more reliable than our postpunishment mechanisms.

1 Introduction

A person commits a crime. The authorities detect evidence of this. It is determined beyond a reasonable doubt that the person committed the crime, and they are punished. This is the ideal operation of postpunishment, of which we are familiar. Now consider another possibility. The authorities detect evidence of an impending crime, and it is determined beyond a reasonable doubt that the crime will be committed (if no interruption occurs). The person is punished. This is prepunishment.

The existing philosophical literature on prepunishment assumes i) that prepunishment exists only in science fiction, and ii) that prepunishment is unproblematic for consequentialists. I will show that both these assumptions are false, and that this calls for changes to our thinking about the philosophical importance and moral status of prepunishment. Prepunishment is morally suspect for both nonconsequentialist and consequentialist reasons and it is not science-fictional: it is a feature of our legal systems as they operate today and its use is increasing alongside our technological abilities. Thus, theorizing about prepunishment is a critical topic of inquiry for moral philosophers and legal theorists.

In Section 2, I review the philosophical debate over the moral status of prepunishment and argue that prepunishment already occurs in real life. If prepunishment is morally wrong,
as nonconsequentialist moral philosophers have argued, then this calls for reforms to current legal theory. Then, in Section 3, I reveal a previously unnoticed problem for prepunishment: according to standard decision theory, prepunishment mechanisms have no deterrent power. This is a serious problem since, from many perspectives, but especially from a consequentialist one, a primary purpose of punishment is deterrence. Finally, in Section 4, I discuss the conditions under which prepunishment systems are likely to generate bad societal consequences. This discussion may help to guide the design of near-future legal systems as technological advancements allow for greater reliance on prepunishment mechanisms.

2 Prepunishment in the Real World

In 1992, Christopher New published a provocative article titled “Time and Punishment.” He began by introducing a thought experiment in which the traffic police know that a person is going to exceed the speed limit tomorrow and decide to write a traffic ticket for the offense today. New claimed, i) that most people will judge this to be morally wrong, and ii) that this intuition is due to “mere prejudice.” The intuition against prepunishment, New [36–7] claimed, results from “attaching improper moral significance to an insignificant temporal fact.” He did grant that the intuition may result from prepunishment being impossible. Nevertheless, he stated that even if no one will ever be in a position to prepunish, “it is coherent to suppose that someone might.”

New’s claims garnered a significant number of responses. While most authors disagreed that prepunishment is morally permissible, all accepted his assumption that prepunishment is currently impossible. Statman (1997, 133) wrote, “The possibility of prepunishment … is a purely theoretical one which, even if accepted, would have no implications for our human institution of punishment.” Indeed, discussions of prepunishment in the philosophical literature add unrealistic assumptions, as in Smilansky (2007, 347): “Let us assume for the sake of our argument both determinism and complete predictability: if people’s actions are determined, and we have perfect epistemic capacities, we can know ahead who will commit a crime.” The only potential real-world application of prepunishment, according to Lloyd Strickland, is a divine one. Strickland (2011, 108) writes, “Needless to say, it is highly unlikely that humans will ever develop sufficient foresight to make prepunishment a genuine option, with prescience seemingly forever to remain the preserve of God.”

In this way, the philosophical literature on prepunishment mirrors the fictional literature. Many philosophers notice the similarity between New’s case and Philip K. Dick’s short story “Minority Report,” which is set in a fantastical future where prepunishments are doled out on the predictions of a trio of “precog mutants.” In the film adaptation, “precrime” agents tap into the minds of these mutants to see visual representations of their predictions. This information is then provided to John Anderton (played by Tom Cruise), who uses impressive futuristic gadgets to apprehend the perpetrator before the crime occurs. Given this backdrop, it is perhaps no surprise that people have assumed that prepunishment is futuristic.

However, prepunishment, or at least the part of it that matters morally, is not futuristic. Here is a story of prepunishment following the pattern of “Minority Report.”

*Science Fiction*

Precogs (mutants who reliably predict the future) report that John is plotting
to murder his former wife, Joan, and will do so soon unless police intervene.
Detectives collect John at his apartment and take him to jail for the crime of
“future murder.”

This story seems fantastic and irrelevant to our world. But now consider the story of John
as it might occur in the real world.

_Real World_
John’s friends report that John is plotting to murder his ex-wife, Joan, and will
do so soon unless the police intervene. Detectives decide to follow John as he
drives to Joan’s house. As he approaches the house, gun in hand, the detectives
arrest John and take him to jail for the crime of attempted murder.

_Real World_ is not fantastic; it has happened many times. What are the differences between
the stories? There are two differences that stand out. The first is that, in _Science Fiction_,
evidence is provided by mutants rather than by John’s friends and the observations of de-
tectives. The second difference is that the crime is called “future murder,” whereas in _Real
World_ it is called “attempted murder.” Both of these differences are morally irrelevant.

First, it is simply a mistake to call the crime “future murder” because no murder ever
takes place: John was about to murder his wife but then he was stopped by the police. The
crime in _Science Fiction_ is better described as “attempted murder.” In this case, the precogs
are predicting what would happen if things were to play out without police intervention
(they are not “seeing the future,” because what they see never occurs). In _Real World_,
the detectives make the same prediction: “what would happen if we do not intervene?”
Accordingly, in each instance we might call the crime either “attempted murder” or, perhaps
more accurately, “counterfactual murder”: if the police had not stopped John, then he would
have murdered his wife.

Second, while it is true that the _source_ of the evidence differs between _Science Fiction_
and _Real World_, it is only the _quality_ of the evidence that matters. _Science Fiction_ features
evidence from mutants. _Real World_ features evidence from John’s friends and the observa-
tions of detectives. Mutants, testimony, and detective observations all provide evidence, in
their respective worlds, for the same proposition: John will murder his wife unless the police
intervene.

There is an ambiguity in our use of the term “attempted murder.” Sometimes it refers
to situations in which someone assaults someone else, with the intent to kill, but the assault
fails to kill. Other times “attempted murder” refers to situations in which someone would
have assaulted someone else, with the intent to kill, but is apprehended beforehand. This
latter concept is the focus of _Science Fiction_ and _Real World_, and the thing most correctly
called “counterfactual murder.” Perhaps, to be most precise, we could call it “counterfactual
murderous assault.”

Let’s take a closer look at current legal theory concerning attempted murder in the United
States. Attempted murder is categorized as an “inchoate crime,” along with conspiracy and
solicitation. Inchoate crimes _anticipate_ a further criminal act.
In most jurisdictions, attempted murder requires an intention to kill and that the offender took some “substantial step” toward the killing. *Model Penal Code*\(^1\) Section 5.01(2) states that “substantial steps” may include any of the following (among others):

a. “Enticing or seeking to entice the contemplated victim of the crime to go to the place contemplated for its commission.”

b. “Lying in wait, searching for or following the contemplated victim of the crime.”

c. “Possession of materials to be employed in the commission of the crime, which ... can serve no lawful purpose of the actor under the circumstances.”

d. “Unlawful entry of a structure, vehicle or enclosure in which it is contemplated that the crime will be committed.”

Given that the *MPC* requires an intention to kill and a substantial step, it’s tempting to think that in *Real World* we’re punishing for having an intention and taking a step, whereas in *Science Fiction* they’re punishing for counterfactual murder. Here is a way to test this hypothesis. What happens when an agent intends to murder and takes a substantial step but then changes their mind? If the *MPC* aims to punish for the intention and step, and not the counterfactual, then we would expect it to recommend that such an agent be found guilty of attempted murder even though they chose to abandon the plan. If, however, the *MPC* is punishing for the relevant counterfactual and not the intention and step themselves, then we would expect an agent who abandons their plan to be considered innocent. The latter is exactly what we find. *MPC* Section 5.01(4) states that an agent cannot be convicted of attempted murder if he “abandoned his effort to commit the crime or otherwise prevented its commission, under circumstances manifesting a complete and voluntary renunciation of his criminal purpose.” This shows that what the *MPC* cares about in an attempted murder charge is not the intentions and steps in themselves.\(^2\)

Intentions and steps are relevant because they provide strong evidence of the counterfactuals: since people often do what they intend to do and they’re more likely to do it if they’ve taken a significant step toward it. These considerations convince me that what the *MPC* cares about most in situations like *Real World* is not a person’s intentions or steps, but the relevant counterfactuals. I find myself in agreement with Helen Beebee, who sums this point up nicely, writing: “The requirement that there be actual intentions and preparations can be seen merely as a reflection of the fact that these are the routes by which prosecutors come to have reasonable beliefs about what will happen if they do not intervene.” This is why *MPC* labels attempted murder an “inchoate crime”: the punishment of such crimes anticipates a further criminal act.

---

\(^1\)The American Model Penal Code is a comprehensive synthesis of American law which states often consult in creating their individual penal codes.

\(^2\) *MPC* 5.01(4) states that the renunciation of criminal purpose cannot be due to the person encountering unforeseen circumstances that made the crime difficult to accomplish or difficult to accomplish without being detected. This shows that the relevant counterfactual has several parts: it is not simply that the agent would have committed the crime had they not been stopped by the police, but additionally that the agent would have committed the crime had they not encountered unforeseen circumstances of this type.
As Yaffee (2014, 142) notes, not all US jurisdictions follow the MPC in allowing for the abandonment defense. However, my argument only requires that some jurisdictions do, since I am arguing that prepunishment sometimes occurs in real life. The view I’m putting forward is that, at least in some real-world cases, an intention and significant step is required for attempt charges because that is a seemingly clear and objective standard by which to make the judgment that a person would have done it. And abandonment is at least sometimes seen as a positive defense for attempt charges because that is a seemingly clear and objective standard by which to make the judgment that a person would not have done it.\(^3\)

If this is correct, then the significance of “Minority Report” has been underestimated. Dick is not merely providing a story about a futuristic world; he’s inviting us to see in a new light some of the workings of our law enforcement practices as they actually exist today. Dick imagines bizarre “precogs” that provide more reliable evidence than detective observation ever could, but he leaves untouched all the features of the situation that are morally relevant. And the interesting thing is that many philosophers have thought that what Dick describes is morally wrong.

As we saw above, New claimed that resistance to prepunishment is the result of “mere prejudice.” Since New’s article, however, a seeming consensus has formed that prepunishment is immoral. In his response to New, Smilansky (1994, 51–2) writes:

> While in postpunishment the offender cannot take back her actions, in prepunishment she still has time to choose. She can decide, even in the last minute, not to commit the offence.... This explanation [for why prepunishment is wrong], it seems to me, is that in prepunishment we are not showing the respect due to the moral personality of the agent, who is, when ‘punished’, as yet innocent, and who we must respect as capable of not committing the offence. In prepunishment there is categorically still time, a ‘window of moral opportunity’ for the would-be offender.

In light of the discussion above, it is clear that Smilansky is not merely providing an “explanation” for something people already believe. Rather, he is committing himself to the view that the abandonment defense does not go far enough. According to Smilansky, not only must we refrain from punishing people if they actually abandon their intention, we must also refrain from punishing them if there is any time left in which they could abandon their intention — even if we know that they’re not going to. Therefore, if Smilansky is right, reforms are required to US legal theory concerning attempts.\(^4\) In fact, the MPC Section

---

\(^3\)Yaffee’s (2010) view of attempts diverges from the MPC in rejecting the abandonment defense (he claims that abandonment should instead be viewed as a mitigating factor at sentencing). However, Yaffee explicitly focuses on counterfactuals in justifying attempt charges. For example, he maintains that an attempt is complete when it is the case that a person would commit the crime if they had the means and opportunity to do so. Therefore, according to Yaffee, an attempt has taken place even if the agent abandons their plan. Yaffee’s counterfactual view is consistent with my counterfactual interpretation of the MPC, but it adds an extra condition to the relevant counterfactual (see footnote 2). According to Yaffee, the relevant counterfactual is that an agent would have committed the crime had they not been stopped the police, encountered unforeseen circumstances, or changed their mind.

\(^4\)Indeed, the reforms motivated by Smilansky’s view seem to be in line with those suggested by Alexander et al. (2009, 197).
5.05(1) goes so far as to recommend the same punishment for attempt as for the consummated offense. In doing so, the MPC doesn’t seem to care about the “window of moral opportunity” whatsoever.5

Other arguments against prepunishment focus on the relationship between free action and prediction. Fred Feldman (1995, 77) writes:

Consider a typical case in which it seems quite likely that a certain person will commit some crime. We think he will deserve the legally mandated punishment only if he will be responsible for the crime and we think that he will be responsible for the crime only if he will commit it ‘freely’; and we think that if he will commit it ‘freely’, then it cannot yet be quite certain that he will commit it. There must still be some possibility that he will decide not to commit it. So we insist upon a legal system that prohibits punishment-in-advance.

As with Smilansky, Feldman is mistaken to think that he is explaining something people already believe. In fact, the US legal system, at least in some jurisdictions, does not prohibit the kind of punishment he identifies (at least, not for murderous assaults we are nearly certain would have been freely done had they not been interrupted). If Feldman is right, then the US legal system is wrong.

As a final example, consider Roy Sorenson’s argument against prepunishment: “My thesis is that a crime justifies a verdict only by being a cause of that verdict. The asymmetry of ‘A causes B’ explains why we cannot prepunish wrongdoers. In practice, causal asymmetry matches temporal asymmetry. This accounts for the appearance that temporal prepunishment is unjust.” Again, this argument cannot be merely an account of existing beliefs. If the unconsummated portion of an inchoate crime cannot be the partial cause of a verdict, as Sorenson claims, then counterfactuals about what would have occurred had the police not intervened cannot help justify attempted-murder verdicts. But, currently, such counterfactuals do help justify attempted-murder verdicts.

I won’t take a stand on whether these arguments against prepunishment succeed. Perhaps the arguments do succeed.6 If so, significant reforms to legal theory are required. In any event, theorizing about the moral status of prepunishment is of direct relevance to the real world. Contrary to what philosophers have thought, a lot is at stake for our real-world punishment practices.

3 The Consequentialist Problem for Prepunishment

Your company, MegaTech, has dispensed with its payroll staff and now leaves its payroll system completely up to the management of individual employees. At the end of the week,

5It’s worth noting that this feature is not unique to US law; many law systems recommend that attempts be punished nearly as seriously as consummated offenses. For example, the UK standard reference text Blackstone’s Criminal Practice states, “Even in cases where a low level of injury (or no injury) has been caused, an offence of attempted murder will be extremely serious” [495].

6This view is foreshadowed in Dick’s “Minority Report,” when Anderton explains, “We’re talking about individuals who have broken no law ... because we get them first, before they commit an act of violence.... In a sense, they are innocent” (Dick, 1987, 72).
you are given access to the vaults so that you can take your salary. To keep employees from stealing, predictions are made on Monday about whether an employee would steal on Friday. These predictions have been found to be reliable enough to establish beyond a reasonable doubt that an employee would steal. Any predicted thieves are fired and sent to jail.

Now consider two ways the story might continue.

**Counterfactual**
It’s Friday. On Monday you were not fired and sent to jail. You are alone in MegaTech’s vaults. It occurs to you that you could take two times your salary, and no one would notice. You are certain that you will not be punished, since any punishment would have occurred on Monday. So, you decide to take the money and, of course, you are not caught.

**Actual**
On Monday you are informed that you have been fired and the police are coming to take you to jail. “Why?” you protest. “I wasn’t planning to take any money! Even if I want the money, I know about the risk of imprisonment, and it’s not worth it to me!” Your boss replies, “Ah, but you see, the predictors saw that you were going to take the money, for consider Counterfactual.”

**Counterfactual** is a description of the closest possible world to Actual in which you are not sent to jail. On Monday, the MegaTech predictors ask, “would you steal if we don’t intervene? Given Counterfactual, the answer is “yes.” And so you are fired on Monday.

Of course, if MegaTech might postpunish you, then you would have a reason on Friday not to steal. However, with a prepunishment system, there is no deterrence on Friday. Imagine that you’re an ideal candidate for deterrence. While in the vault on Friday, you might explain: “If there were any risk whatsoever that taking this money would cause a punishment, I would not take this money. In other words, I am maximally deter-able. But there is no such risk; nothing I do now can change the past.” Thus, MegaTech’s prepunishment policy would have no deterrent power whatsoever. Even an extraordinarily ineffective postpunishment system, which detected criminals only a small percentage of the time, would have more deterrent power.

The case of MegaTech follows the pattern of the cases presented in “Minority Report”: the punishments prevent the crimes and, in some instances, the agents do not form the relevant intention before they are convicted. However, some proponents of prepunishment have argued that it is only justified given two elements: i) the agent actually forms the intention to commit the crime, and ii) the punishment does not prevent the crime. No matter: a fascinating feature of the deterrence problem is that it remains whether or not agents form an intention to commit, or actually do commit, the predicted crime.

Consider a variant on the case of MegaTech in which the prepunishment is a dozen lashes, to be administered on Monday, and after which you free to act as you please. Even if you receive the lashes on Monday, you are subsequently put in a decision situation in which you know that you could take more than your salary without causing any (further) punishment. Thus, whether prepunishment is preventative or non-preventative is irrelevant to its deterrence problem.

Whether or not intentions are required has no effect on deterrence either. Regardless of whether you intended on Monday to steal the money on Friday, on Friday you will still
know that you can take more than your salary without causing punishment. At that point, the prepunishment system is irrelevant, because you can’t change the past. Supposing that prepunishment, whether it requires an intention or not, does not deter you.8

Consequentialists about punishment, such as utilitarians, think that punishment should aim to produce good consequences. One of the primary ways punishment produces good consequences is by deterring criminals. However, all of the arguments against the moral permissibility of prepunishment in the philosophical literature are nonconsequentialist and do not reference deterrence. In fact, they assume that the deterrent value of prepunishment is the same as, or greater than, the deterrent value of postpunishment. This assumption starts with New (1992, 38): “In pre- as much as in postpunishment the penalty imposed may deter potential offenders, as also the actual offender, from committing other offences in the future.” In his influential reply, Smilansky (1994, 50) agrees with New at least to that extent. He writes, “We need not concern ourselves here with [consequentialist objections]. I allow prepunishment may be useful... The crucial question is whether someone who believes in nonconsequentialist constraints of justice has the resources to reject [prepunishment].”

---

7 The intention-required version of the case mirrors that of Kavka’s (1983) famous Toxin Puzzle, in which you are offered one million dollars to intend today to drink a toxin tomorrow. As Kavka argues, the proposal only gives you a reason to try to get yourself to intend to drink the toxin today; it does not give you a reason to actually drink the toxin tomorrow. The same is true in the intention-required version of the MegaTech case. MegaTech’s policy would give you a reason to try to refrain, somehow, from forming the intention to steal on Friday, but it would not give you a reason to refrain from stealing on Friday.

8 Decision-theory enthusiasts may notice a similarity between cases of prepunishment and Newcomb’s problem. In fact, there is a close relationship between prepunishment decision situations and the various types of Newcomb problems that feature in philosophical decision theory.

Non-preventative prepunishment creates decision situations that resemble the “transparent” Newcomb problem (Gibbard and Harper, 1981, 181–2), in which the money in the boxes is revealed before an agent decides whether to one- or two-box. Both evidential and causal decision theory recommend two-boxing in the transparent Newcomb problem, and consequently they hold that non-preventative prepunishment does not deter. For example, the variant of the MegaTech case in which the penalty is a dozen lashes is a transparent Newcomb problem in which stealing money is equivalent to two-boxing. Either it has been predicted that you would steal or it has been predicted that you would not. In either case, both evidential and causal decision theory agree that you should not regard potential past punishment as a reason to not steal now.

Preventative prepunishment is a variant of the transparent Newcomb problem in which an agent is prevented from having the choice to one- or two-box if it is predicted that they will two-box. This is sometimes called a “Parfit’s Hitchhiker” in honor of Derek Parfit (1984, 7)’s influential case. As with transparent Newcomb problems, both evidential and causal decision theory recommend two-boxing in Parfit’s Hitchhiker, and consequently they hold that preventative prepunishment does not deter. For example, the original case of MegaTech is a Parfit’s Hitchhiker. You are only allowed into the vault if it is predicted that you would choose to not steal. But the prior possibility of prepunishment is, at that point in time, irrelevant. Evidential and causal decision theory therefore agree: if you are allowed to enter the vault, then a prepunishment system should not deter you from stealing.

Thus, according to both causal and evidential decision theory, which make up what philosophers consider to be standard decision theory, the cases of prepunishment presented in “Minority Report” and the philosophical literature have no deterrent power for rational agents. These cases are either transparent Newcomb problems or Parfit’s Hitchhikers, and rationality requires two-boxing in these cases.

There are some proposed decision theories that do recommend one-boxing in the transparent Newcomb problem and Parfit’s Hitchhiker. See Gauthier 1986, Chapter 6; McClennen, 1990; Meacham, 2010; Greene, 2018; Soares and Yudkowsky, 2018; and Soares and Levinstein, manuscript. According to these accounts of rational decision making, prepunishment systems do deter rational agents. However, these accounts are currently viewed as nonstandard.
More recent discussions of prepunishment emphasize the supposed deterrent value of prepunishment. Statman (1997, 129) writes, “Utilitarian benefits that can be gained by punishing after crime (‘postpunishment’) can also be gained by prepunishment, e.g., deterring other potential criminals…. It is a puzzle mainly for retributivists.” Petersen (2014, 140) writes:

According to a utilitarian rationale, the purpose of punishment is crime reduction (by means of strategies like rehabilitation, incapacitation, or deterrence). And if, for example, a neuroprediction based on a brain scan of an offender X’s future behavior tells us that X is likely to commit new crimes in the future, there is a utilitarian reason to punish X more harshly than if X were predicted not to be dangerous.

Surprisingly, and on the contrary, the case of MegaTech reveals that an important problem with prepunishment stems from consequentialist considerations, and it is a problem particularly for deterrence theorists. The problem is that prepunishment has no deterrent power.

If prepunishment does not deter, what does it do? The difference between the effects of post- and pre-punishment is the difference between deterrence and character control. Before turning to character control, let us take a closer look at the nature of deterrence.

There is no universally-accepted and unproblematic definition of deterrence. Clearly, deterrence has something to do with discouraging an action based on fear of the consequences. However, it has not been formally specified, in either a philosophical or legal context, when the discouragement must occur and whether it must be discouragement of the action directly or of related intentions and other dispositions. I propose that we identify deterrence with direct discouragement of an action at the time of decision. Therefore:

Punishment systems deter if they give agents reasons not to perform certain actions at the time of decision through influencing a person’s assessment of the probability of outcomes.

When we create a system of postpunishments for the purpose of deterrence, we assume that agents prefer not to undergo the punishment, and that our system raises the probability of punishment conditional on the person committing the crime.

On the other hand:

Punishment systems control character if they give agents reasons to change their preferences over outcomes or to self-bind.

If your company adopted MegaTech’s prepunishment system, you’d constantly have to ask yourself questions like this: Would I prefer to steal money if I knew I could get away with it? Or, am I disposed to steal money in a situation in which I know I can get away with it? If the answer is yes, then the police will come for you on Monday. You would need to start figuring out a way to change your preferences, or to self-bind yourself to relinquish rational control of your future decisions, to ensure that you would not steal even if you knew you could get away with it. These effects of prepunishment I identify with character control.
One justification for these definitions is that what I call “deterrence” is the thing that deterrence theorists have claimed is an obvious benefit to society. And the things that fall under “character control” are not things that deterrence theorists have claimed are obvious benefits.

4 Effects of Prepunishment Systems on Society

Consider three ways an agent might respond to the threat of legal punishment. First, willing compliance, in which an agent complies with a law and would do so without the threat of punishment. Second, grudging compliance, in which an agent complies with a law but would not do so if there were no threat of punishment. The third option is noncompliance.

The deterrent aspects of postpunishment push toward grudging compliance, while the character control aspects of prepunishment push toward willing compliance. At first glance, then, it might seem that prepunishment is desirable to postpunishment, given the plausible assumption that willing compliance is preferable to grudging compliance (which is, in turn, preferable to noncompliance). However, the transition from grudging compliance to willing compliance does not justify all invasions of personal liberty. In fact, not even the transition from noncompliance to grudging compliance seems to provide a universal justification: Western criminal justice systems do not see themselves as forcing, under threat, a change in criminals’ preferences over outcomes, even if this would remove their noncompliance.

The desirability of character control depends on large issues of political philosophy concerning the justification of punishment. From some perspectives in political philosophy, the considerations raised above are enough to determine that prepunishment is unjust or unfair, regardless of its consequences. From others, prepunishment is, at least in principle, permissible. For the purposes of this paper, I will focus on a narrower topic: the expected consequences of different systems of prepunishment on groups of rational agents trying to maximize their expected utility. This discussion will certainly help inform, if not determine, the desirability of prepunishment systems in society.

4.1 Pure Prepunishment Systems

In a pure prepunishment system, like that presented in “Minority Report,” prepunishment has completely replaced postpunishment, and thus there are no postpunishment mechanisms whatsoever. A pure prepunishment system can be viewed, as it is in “Minority Report,” as the natural culmination of a legal system as its prepunishment abilities increase. After all, what need is there for postpunishment mechanisms once a system’s prepunishment mechanisms are reliable enough to catch all crimes before they occur? Pure prepunishment systems...
Figure 1: Decision Trees for Pure Prepunishment Situations
are unlikely to be developed in the near future, but they provide a useful backdrop for evaluating the consequences of mixed systems.\(^\text{10}\)

Pure prepunishment systems represent a serious hazard to the grudgingly compliant. The hazard is created by the lack of postpunishment mechanisms: if there is no postpunishment, then people are constantly put in situations where they could break the law and get away with it. This lack of postpunishment makes no difference to the willingly compliant — they are willing to keep within the law even if they could get away with it. It also makes no difference to the noncompliant — they are not deterred by postpunishment in the first place. However, for the grudgingly compliant the difference is crucial: whereas in a postpunishment system the grudgingly compliant decide to keep within the law for fear of the consequences, in a pure prepunishment system they are sent to jail before they even have a chance to decide.

The hazard for the grudgingly compliant was on display in the original MegaTech case. MegaTech’s pure prepunishment policy will not succeed in deterring the grudgingly compliant; instead, the policy will ensure that only the willingly compliant work at MegaTech. Imagine that you are grudgingly compliant and consider the decision tree representing the MegaTech case in Figure 2a. Consider your decision situation at node \(d_1\): you can either steal money from the vault or take your normal salary. Since you are only grudgingly compliant, and at this point there is no chance of punishment, you would choose to steal. However, before this, MegaTech’s predictors predict whether you will steal at \(d_1\). If they predict that you would steal at \(d_1\), then you are sent to jail. Therefore, you never reach \(d_1\), and instead end up in jail.

Extrapolating from this, we can expect the results of MegaTech’s policy to be different than what its executives might have intended. Shifting from a postpunishment to a prepunishment policy results in a purge of all grudgingly compliant employees who were previously deterred by the postpunishment mechanisms. Something similar is true at the level of society: enacting a pure prepunishment system would tend to drastically increase the number of people sent to jail because it would convict both the non-compliant and the grudgingly compliant. It would, in effect, result in a mass purge of all but the willingly compliant.

People disagree about the point of punishment: is it to prevent future crimes, to give the guilty what they deserve, to internalize society’s norms, or something else? Whatever the case, the point of punishment is not to round up all the people that would commit crimes if they could get away with it. But that is what a pure prepunishment system would do.

Perhaps, however, there are ways for the grudgingly compliant to avoid jail without becoming willingly compliant. If MegaTech forewarns employees of the new policy before it is enacted, then employees would have a chance to quit the job before the first predictions are made. This situation is represented by \(2b\), where you have the opportunity to decide at \(d_0\) to quit before the prediction is made. The utility of this outcome (0) is less good than those available at \(d_1\) (1 and 2), but you do not expect to reach \(d_1\).

Consider a corresponding case involving the prepunishment of murder. John, who lives in a pure prepunishment system, would murder his ex-wife, Joan, if he could get away with it. However, John knows that he would only kill Joan if he had a gun, which he currently

\(^{10}\)In this section, I focus on preventative prepunishment for two reasons. First, actual prepunishment practices currently work in a preventative way, and I believe that this will continue to be the case as prepunishment mechanisms increase. Second, as I argued in Section 3, the difference between preventative and non-preventative prepunishment has no effect on the deterrent power of prepunishment.
does not possess. Does a pure prepunishment system deter John from purchasing a gun? This depends on when in time the prediction occurs. If the prediction that sends John to jail occurs before he decides whether to buy a gun, then prepunishment would fail to deter him. This is represented by $2c$. In this case, the relevant prediction concerns $d_1$: John’s decision whether to murder Joan once he has a gun. However, the prospect of a prior prepunishment does not give John a reason to refrain from murdering Joan at $d_1$, and thus if he reached $d_1$ he would choose to murder her to attain his highest utility outcome (2). By the same inference, the prospect of prepunishment also does not give John a reason to refrain from buying a gun at $d_0$: nothing John does at that point can change what has already been predicted.

Now consider $2d$, in which the prediction occurs only after the gun is bought. In $2d$, the prospect of prepunishment does deter John from buying a gun. John’s decision at $d_1$ is upstream of the prediction that could send him to jail, and thus what he decides can affect what is predicted. If John never buys a gun, then he would never commit the murder. Thus, John will not be prepunished for the murder if he does not buy a gun.

We can think of each of these strategies for the grudgingly compliant — quitting one’s job in $2b$ and deciding not to buy a gun in $2d$ — as types of self-binding strategies. In the decision theory literature, sometimes self-binding is presented as the ability to take away rational control of one’s future decisions: “self-binding” oneself to not steal or murder is like removing the option to steal or murder at $d_1$. Choosing to quit one’s job or to not buy a gun at $d_0$ has a similar effect: it ensures that the agent will never be in a position to steal or murder, because they will never reach $d_1$. Since these strategies have similar effects, it is useful think of them as representing two ways one might self-bind.

This reasoning reveals an interesting potential non-epistemic justification of the requirement that offenders take a “significant step” before being charged with attempted murder. In Section 2, I suggested that this requirement is best justified by the fact that significant steps are a clear and objective standard by which to make judgments about what a criminal would have done. We can now see that the significant-step requirement also creates a deterrent effect for what would otherwise be a deterrent-less punishment mechanism. If a criminal is not capable of committing a crime until a certain “significant step” is undertaken, and prepunishment predictions are not valid until such a step occurs, then would-be offenders have a reason to not take a significant step. In other words, would-be offenders are deterred from taking a step that puts them in a position to commit a crime (though prepunishment continues to have no deterrent effect on a person who has already put themselves in such a position). If, however, prepunishment predictions are viewed as valid before a significant step is undertaken, then prepunishment has no deterrent effect on significant steps.

Overall, then, the expected consequences of pure prepunishment systems on a society are the following. The move to pure prepunishment makes no difference to people who are willingly compliant or non-compliant, but it does represent a significant hazard for the grudgingly compliant. To avoid prepunishment, the grudgingly compliant either need to change their preferences to become willingly compliant or they need to engage in self-binding strategies that remove their ability to act illegally. Self-binding strategies are easier to implement if prepunishment predictions are not valid until a “significant step” is taken. If most of the grudgingly compliant are unable to become willingly compliant or self-bind, then pure prepunishment systems could result in their mass incarceration.
## 4.2 Mixed Systems

Mixed systems combine both pre- and postpunishment mechanisms. As with pure prepunishment systems, the effects of a mixed system are relevant only to the grudgingly compliant. The crucial variable is the difference in reliability between the prepunishment and postpunishment mechanisms. For any given crime, as the prepunishment reliability increases and the postpunishment reliability decreases, the hazard for the grudgingly compliant becomes worse (the hazard is most extreme in pure prepunishment systems). Conversely, as the prepunishment reliability decreases and the postpunishment reliability increases, the hazard for the grudgingly compliant diminishes (the hazard is eliminated in pure postpunishment systems).

To see this, first notice that the hazard for the grudgingly compliant is relatively low when both pre- and postpunishment mechanisms are each 50% reliable in detecting crime. Imagine a modified MegaTech case in which this is so. Recall that since you are grudgingly compliant, you would steal money from the MegaTech vault on Friday if you knew you could get away with it. However, there is a 50% chance of postpunishment should you steal money, and so you are deterred from stealing. Since you would not steal money on Friday, you are not prepunished on Monday.

This case is represented by the first row of Figure 3. There is a 50% chance of prepunishment and a 50% chance of postpunishment, which results in a 75% overall chance of some form of punishment for people who are willing to risk stealing money on Friday. However, by choosing to steal on Friday, a person is only risking a 50% chance of punishment, since, on Friday, the potential for prepunishment is irrelevant. Nevertheless, the difference between a 75% or 50% chance of punishment should usually make little difference.

Now consider a situation more heavily weighted toward prepunishment: a 90% chance of prepunishment and a 10% chance of postpunishment. This results in the overall chance of punishment being 91% for anyone who is willing to risk stealing money on Friday. However, such a person would only be taking a 10% risk of punishment by deciding to steal on Friday. Thus, there is a serious hazard for the grudgingly complaint. A person may be willing to risk a 10% chance of punishment but unwilling to risk a 91% chance of punishment, but the mixed pre- and postpunishment mechanisms force a 91% chance of punishment on anyone willing to risk a 10% chance. As with pure prepunishment, to avoid this hazard such a person needs to either change their preferences (so that they are no longer willing to risk a 10% chance of punishment) or engage in self-binding strategies.
When the chance of postpunishment is significantly greater than the chance of prepunishment, the hazard for the grudgingly compliant virtually disappears. For example, if there is only a 10% chance of prepunishment but a 90% chance of postpunishment, a person who steals money is willing to risk a 90% chance of postpunishment and faces a 91% chance of punishment overall. This should make little difference.

Modern Western punishment systems are like this last situation. Only a subset of crimes allow for attempt offenses, and only a subset of attempt offenses are instances of prepunishment. Even in cases where a crime allows for both pre- and postpunishment, prepunishment is usually more difficult to enforce. In modern Western legal systems, then, it is no surprise that the deterrence problem with prepunishment and its hazard for the grudgingly compliant has gone mostly unnoticed. The hazard is created when prepunishment mechanisms are more reliable than postpunishment mechanisms, and in modern systems, the reverse is true: postpunishment mechanisms are currently dominant.

5 Conclusion

I have argued that prepunishment is a real feature of modern legal systems. Therefore, if prepunishment is morally impermissible for nonconsequentialist reasons, as philosophers have argued, then modern legal systems should be reformed. Further, I have argued that prepunishment is deeply problematic for consequentialist reasons. Most importantly, prepunishments do not deter crimes but instead compel would-be criminals to try to change their preferences over outcomes or to self-bind. The negative consequences of prepunishment are most obvious in pure prepunishment systems and in mixed systems in which the prepunishment mechanisms are more reliable than the postpunishment mechanisms. Modern legal systems are mixed systems in which the postpunishment mechanisms are dominant, and thus the consequentialist problems with prepunishment are hidden. However, the importance of the models presented in this paper will increase as technological and societal advancements improve our ability to prepunish. These models suggest that it will prove crucial to ensure that the reliability of a system’s prepunishment mechanisms never outrun the reliability of its postpunishment mechanisms.

References


