Res Publica (2022) 28:557–583
https://doi.org/10.1007/s11158-022-09567-y

PG ESSAY PRIZE

Entrapment and Manipulation

Jonas Haeg[1]

Accepted: 29 September 2022 / Published online: 21 November 2022

Abstract

Why is wrong to punish criminals who have been entrapped by the state? The paper begins by presenting some criticisms of existing answers to this question. First, they fail to put the target, or victim, of entrapment at the centre of the moral explanation. Second, they fail to account for the intuitive relation between the reasons not to entrap and the reasons not to punish. Third, they struggle to account for the existence of agent-neutral reasons not to punish entrapped offenders. Lastly, they are ill-equipped to explain why entrapment seems problematic also outside the legal context. In response, the paper develops a novel account of entrapment: the Manipulation Account. According to this, entrapment always involves a particular kind of manipulation (manipulation-by-hidden-intentions) which morally taints punishment. In short, I suggest that both the initial entrapment and the subsequent punishment involve wrongful manipulation. Lastly, the paper presents some untraditional, but ultimately welcoming, implications of the account.

Keywords Entrapment · Manipulation · Punishment · Police ethics

Earlier versions of this paper were presented, and benefitted from the audience's feedback, at the Oxford Graduate Conference in Political Theory (2020), the Zagreb Applied Ethics Conference (2021), the Rocky Mountain Ethics Congress (2021) and the London Graduate Workshop in Moral and Political Philosophy (2022). For written comments and helpful discussions, I am particularly grateful to Romy Eskens, Sarah Fine, Connor Kianpour, David Owens and Massimo Renzo. The paper was also greatly improved by constructive comments and suggestions from the judges of the *Res Publica* Postgraduate Essay Prize, as well as an anonymous reviewer of this journal.

✉ Jonas Haeg
Jonas.heen.haeg@gmail.com

[1] Philosophy, King's College London, London, UK

Introduction

Entrapment occurs when police officers, or agents of the state, solicit, persuade or otherwise encourage a person to commit a crime they would not otherwise have committed, for the purpose of arresting and punishing that person. As an example, consider the following case:

> *Drugs*: An undercover police officer, P, is posing as a drug dealer at a university party. When he sees D, he asks D to deliver some drugs to a house down the street. D declines but P is persistent, saying he really needs the sale but is too busy working at the party. Ultimately, D agrees. He is later arrested by other police officers when he arrives at the house with the drugs.

The topic of entrapment is puzzling for the following reason. On the one hand, most people share the intuition that it is wrong to entrap D and to subsequently punish D. (At the very least, D ought to get a reduced sentence). After all, he would not have committed the crime if it had not been for P pushing him to do so and P should not have pushed him to do it. On the other hand, many also hold that D seems to satisfy the criteria for being fully liable to punishment. He committed a crime without justification or excuse. The fact that someone was persuaded into committing a crime does not ordinarily amount to an excuse (or justification). Compare D to his counterpart who was persuaded by a *genuine* drug dealer. D appears just as culpable for the crime as this counterpart would be and, so, appears to be liable to the same amount of punishment. Indeed, in D's mind he *is* the counterpart.

These two conflicting intuitions give rise to what we may call 'the puzzle of entrapment'. In short: although it seems wrong to punish D, he seems just as deserving of, or liable to, punishment as his non-entrapped counterpart would be. The aim of this paper is to resolve this apparent tension in a novel way. According to what I call the Manipulation Account, entrapment always involves a form of manipulation which makes it *pro tanto* wrong to punish targets of entrapment. More precisely, I argue that entrapment involves a particular kind of manipulation (manipulation-by-hidden-intentions) which is not necessarily culpability-affecting. That is why D is liable to or deserving of full punishment. However, this form of manipulation 'morally taints' the punishment of otherwise liable offenders, such as D. The reason is that to punish entrapped offenders is to fulfil or complete the wrongful manipulation in question and we have *pro tanto* duties to avoid completing wrongs such as manipulation. Thus, it is *pro tanto* wrong to punish entrapped offenders.

I say 'pro tanto' wrong because, in my view, entrapment does not always make punishment impermissible. As I discuss more in detail below, there can be cases in which it is all-things-considered permissible to punish an entrapped offender. For instance, it may be permissible to punish targets of entrapment who have been radicalized by undercover agents and pose a serious danger to society if not punished for their crime. Or, if the crime in question is sufficiently grave and harmed innocent people, then the state may owe it to the victims to hold the offender to

account. But in all these scenarios, we should nevertheless hold that the targets have suffered a wronging by being entrapped and will be wronged by being punished. It is just that the state has a *lesser evil* justification that makes it all-things-considered permissible to inflict this wrong.[1] For ease of exposition, I omit the 'pro tanto' qualification in the remainder of the paper.

Understanding what is wrong with entrapment, and how it taints punishment, is not just interesting as it helps solve an intellectual puzzle. It is also important for practical purposes. For instance, it can help us understand where to draw the line between permissible and impermissible pro-active law enforcement, and it can help us see what is good and bad about current legal doctrines concerning entrapment. Towards the end of the paper, I draw out some of the revisionary implications that my account has for current legal views on entrapment.

The paper proceed as follows. 'Problems with Existing Views' motivates the need for a new solution to the puzzle by outlining some general objections to many of the existing views. 'The Concept of Entrapment' argues that entrapment always aims at, or intends for, punishment. Building on that, I develop 'The Manipulation Account' and then defend it against some potential 'Objections'.

Problems with Existing Views

The motivation for the Manipulation Account is that existing accounts of entrapment fail to adequately solve the puzzle of entrapment. Demonstrating that this is the case for all existing proposals is beyond the scope of the paper, but I will outline some quite general worries. Before that, I should note that not everyone believes that entrapment constitutes a puzzle. The puzzle, recall, is generated by two seemingly incompatible intuitions elicited by cases such as *Drugs*:

1. It is wrong to punish (or fully punish) entrapped offenders.
2. Entrapped offenders are liable to (full) punishment.

Some deny the existence of a puzzle by rejecting (1) or (2). Howard (2016) is sceptical of (1). Yet he does not really argue for its rejection, and it seems to me implausible to reject it in paradigmatic cases of entrapment such as *Drugs.* Hughes (2004) and Kim (2019) are both sceptical of (2).[2] They think that entrapment always involves pressuring which is severe enough to diminish the target's responsibility. As such, entrapped agents will always be less than fully responsible for their crimes, which is why they are not liable to full punishment. Although I agree that

[1] There can be cases in which punishment would not even *wrong* the offender. This is the case if the offender has made themselves liable to be entrapped, in which case the entrapment would not be wrong. I return to this in 'Mere Opportunities'. The primary focus of the paper, however, is wrongful entrapment.

[2] It is worth highlighting that Kim has one of the more comprehensive views out there, as he also subscribes to two different accounts of why it is wrong to punish entrapped offenders which do not depend on the claim that entrapment reduces culpability. I return to those below.

entrapment *often* involves severe pressuring, this is not always the case. For instance, D's responsibility does not seem diminished in *Drugs*. To see this, imagine a version of the case in which P is a genuine drug dealer. It does not seem that the mere fact that this dealer persuaded D to commit a crime is sufficient to undermine D's responsibility.[3]

Understandably, then, most authors want to solve, rather than reject, the puzzle. But no one has found a completely compelling solution. Most of the existing accounts suffer from a set of quite general problems, I believe. And it is these problems which motivate the need for an account like the Manipulation Account. To show this, let me first briefly describe a few of the existing accounts.[4]

Consider, first, the Standing Account. It is often said that even if A is worthy of being blamed because they have done something wrong, it can be wrong for certain others, like B, to blame A if B's blaming would be hypocritical because he is guilty of similar wrongdoings, or if B is complicit in A's wrongdoing. In those cases, B is said to lack the *standing to blame* the blameworthy A.[5] Some think that this framework helps explain why it is wrong for the state to punish entrapped offenders as well. Ho (2011), for instance, argues that the state lacks the standing to blame and condemn offenders for crimes which the state has wrongfully instigated or caused—i.e., those which it is complicit in. Similarly, Kim (2019) thinks both hypocrisy and complicity concerns undermine the state's standing to blame or punish when it seeks to blame offenders for crimes it has intentionally created and bears some responsibility for.

Next, consider Kim's Legitimacy Account. [6] He makes the following kind of argument to explain why it is wrong for states to punish entrapped offenders (2019, p. 84):

(1) A precondition for a state's right to punish criminals is that it is committed to *crime prevention*.
(2) A state which entraps people *creates* crimes and therefore violates the precondition for the right to punish.
(3) Therefore, a state which entraps does not possess the right to punish.

In a similar vein, Carlon argues that a precondition for the 'right of prosecution' is that a state embodies the principles of justices but claim that states which entrap

[3] Indeed, it would be troubling if it did. People would be afforded (partial) excuses left and right as long as they could demonstrate that someone else persuaded them into committing a crime. That would mean that only those who motivate themselves to commit crimes would be liable to full punishment. That is obviously not the case.

[4] For a great overview of even more accounts, and some of the problems with them, see Howard (2016).

[5] For more on hypocrisy and lack of standing to blame, see, e.g., Cohen (2006) and Wallace (2010).

[6] To clarify, Kim seems to think that the Legitimacy Account is most important because the Standing Account can only account for why the state lacks standing to *blame* offenders. But many think that punishment is justified on grounds other than blame, so a state could justifiably punish even without having the standing to blame.

fail to satisfy this precondition because '[i]n its unjust desire to punish, the state has ceased to embody the principles of justice' (2007, p. 1123).

Next, consider the Incoherence Account as outlined by Duff et al. (2007).[7] The account is complex, but it boils down to the idea that 'the normative validity of the trial rests on the validity of the state's conduct pre-trial' (2007, p. 236). More precisely, they argue that the state acts incoherently when it seeks to punish someone it has entrapped. The reason is that in entrapping the person the state is—through its encouraging actions—expressing that the criminal behaviour is not worthy of condemnation. However, when a state seeks to punish a criminal for his behaviour, it necessarily needs to express that the behaviour in question is worthy of condemnation. Because such incoherent behaviour undermines the integrity of the trial and the criminal justice system, it is wrong for the state to punish those it has entrapped into committing crimes.

These accounts, as well as others, capture different features that all seem relevant to the explanation of what is wrong with punishing entrapped offenders. But I think they fail to provide adequately complex accounts of this. They miss some important moral features and are therefore both extensionally inadequate (as I argue below) and unable to give us complete and accurate moral explanations even in the cases that they do give us correct verdicts about. Let me start with the second complaint. In looking at many accounts of entrapment, it is easy to feel that people lose sight of, arguably, the most important person and the most important action: the target of the entrapment and the act of entrapment. Both things turn out to do little work in the explanation of why it is wrong for the state to punish someone it has entrapped. Instead, the chief focus is placed on the state itself and the various principles and expectations it is required to live up to.

On the Incoherence Account, for instance, the work is done by the moral importance of upholding the *integrity* of the state, and the fact that incoherent behaviour undermines it. This understanding is reinforced by Andrew Ashworth's version of the view as well. For him, the worry is that a prosecution tainted by incoherent behaviour from the state "would damage the integrity of the criminal justice system" (1999, p. 307). It is thus the importance of the state's integrity which is doing the moral work in explaining why the state should not punish D. D himself, and the fact that he was wrongfully entrapped, play little direct role in this explanation.

Similarly, on the Legitimacy Account, the explanatory work is primarily done by the fact that, in entrapping D, the state simultaneously violates one of the preconditions for having the authority to punish its citizens. It is wrong to punish D, then, primarily because of the state's lack of authority. The fact that D is a victim of wrongful entrapment is largely left out of the explanation. Kim also explicitly accepts this. In outlining the account, for instance, he says that its 'rationale is all about [the state] and not about [the target of the entrapment]' (2019, p.83).[8]

[7] See Dworkin (1985) for an earlier development of this view. Ashworth (1999) also offers some support for a kind of Incoherence Account, in particular highlighting the importance of the integrity of the criminal justice system.

[8] To be clear, Kim thinks that D, as a victim of wrongdoing, is part of the *full* moral picture since, as we have seen, he also believes that entrapment necessarily reduces responsibility. So, part of the explanation of why it is wrong to punish D is also that he is less than fully guilty for the crime. As said, however, I do not think the responsibility-reducing feature of entrapment is essential. In those cases, then, Kim's

This is also true of the Standing Account. In outlining his version of this view, Ho explicitly asserts that according to it, a stay of proceedings in entrapment cases is *not* 'granted to protect the entrapped or uphold any of his or her rights' (2011, p. 95). It is rather granted because the state has lost *its standing* to hold the offender to account for the crime. Again, then, the main claim is that the state has morally compromised itself and its ability to occupy the moral high ground required to blame anyone for the crime in question.

Though this focus on the state itself, and how it can morally compromise itself by creating crimes, captures something morally important about entrapment, I think it is wrong to ignore, in the way they do, the *targets* of entrapment. Intuitively, part of the explanation of why it wrong to punish D is precisely that it *wrongs D* and, moreover, that it wrongs D because he was *entrapped* into committing the crime in question.[9] The duty not to punish D is a duty which is in part *owed to D because he was entrapped*. To see this, notice, for instance, that it seems reasonable for D to feel very resentful towards the state if they punish him. This cannot be so easily captured by accounts that only focus on the state's lack of authority, integrity or standing. For a state can come to lack these things in other ways without it giving the punished person the same kind of complaint. Suppose that X has committed a crime but was not entrapped into doing so. Yet the state has, for instance, recently entrapped others into committing similar crimes. The state may for that reason plausibly lack the integrity, authority or standing to punish X for his crime. But it seems that X's complaint against being punished is quite different and much weaker than D's complaint against being punished. The obvious difference is that D was *entrapped* into committing the crime, but X was not. So, our full account of entrapment should also make this factor morally relevant.

Furthermore, it seems that the wrong of entrapping D and the wrong of punishing D are intimately connected. There is some sort of continuity between the reasons not to punish D and the reasons not to entrap him in the first place.[10] To see this, notice, for instance, that the severity of the wrongness of entrapment appears connected to the prospect of punishment. Suppose that the state has no intention to punish D. The police officers induce him to deliver some drugs because they want to lure out the real drug dealer whom they want to arrest. Now, it may still be wrong to encourage D to break the law.[11] But it seems much less wrong than it would be if punishment was likely to follow. D's complaint against the state's entrapment seems much

Footnote 8 (continued)
full account does not seem able to bring D himself into the explanation of what is wrong about punishing him.

[9] Dillof's (2004) Fairness Account may capture this *wronging* aspect of entrapment. In short, he argues that it is unfair of the state to arbitrarily pick out D to use him for everyone else's benefit (through the general deterrence effect that punishing him could have). Moreover, one is plausibly *wronged* by being treated unfairly by the state. It seems to me, though, that even if that is true, entrapment is particularly problematic because it wrongs people like D for more (serious) reasons than merely the fact that it is unfair. See also Howard (2016, p. 27) for a critique of Dillof's view.

[10] For more on continuity between moral duties, see, e.g., Gardner (2011) and Tadros (2020a).

[11] See, e.g., Tadros (2020b) on why it is bad for people to do wrongful things, and Howard (2016) on why it is wrong to make others more likely to act culpably.

stronger in this latter case. In that sense, the wrongness of punishing D is not like a separate wrong. Rather, it is intimately connected with the initial wrong. Preventing the entrapment from being followed up with punishment is a way of mitigating the severity of the initial wrong. Again, the previously mentioned accounts of entrapment seem to lack the resources to capture this aspect of the moral explanation of why it is wrong to punish targets of entrapment.

The focus on the state, and its having compromised itself, as the main reason for why it is wrong to punish entrapped offenders also means that, in adding to being explanatorily incomplete, these views are extensionally inadequate. The fact that entrapment is a wrong, intimately connected with the likelihood of punishment, can make it wrong for a state to punish an entrapped offender even when punishing them would not threaten its integrity, authority or standing. The reasons not to punish an entrapped offender do not exist only when it is one and the same entity (i.e., the state) which is responsible for the entrapment and the punishment. Yet all the previously discussed accounts make this an essential part of their explanations. On those views, there would seemingly be nothing morally problematic about a state punishing an entrapped offender if another entity was responsible for the entrapment. Since the state itself has not induced the crime, it will not behave incoherently by condemning it now, nor will it have lost its authority or standing to blame the offender. But even though the reasons not to punish an entrapped offender are strongest when the same state is responsible for the entrapment, it is not true that these reasons exist *only* in those cases. Consider, for instance:

> *Treaty:* States A and B have an agreement: any citizen of A found guilty of a crime in B will be punished in A, and vice versa. Unbeknownst to State A, State B entraps a group of people. Among the group members is a citizen of State A.

Despite the agreement, I do not think State A should punish their own citizen, even though he may be fully culpable. Moreover, it should not punish him precisely because he was entrapped into committing the crime. Thus, the fact that he was entrapped seems to provide some reason against punishing him even in cases in which the punishing the state's own incoherence, authority or standing is not at issue.

The point can be illustrated within one state as well. In discussing entrapment, the subject of debate tends to be 'the state' conceived of as one entity which is responsible for both the entrapment and the punishment.[12] In reality of course, the state is composed of many individuals, institutions and agencies. If there is one entity which has (i) directed the police officers to entrap someone, (ii) directed the state attorney to prosecute the entrapped offender and (iii) instructed the court to convict them, then clearly one and the same entity is responsible for both the entrapment and the

[12] An exception is Ho who routinely distinguishes between 'the executive' and 'the court', seeing these as two independent arms of one entity (the state).

punishment. Perhaps we could even say the same in cases in which there are separate entities within the state (police, attorney, court, etc.) responsible for each part but there is collaboration between them. But suppose that a state has done its best to outlaw entrapment practices at all levels, yet some individual police officers still engage in it.[13] It seems morally problematic for the state to punish entrapped offenders even in this scenario, but it is not clear that it is one and the same entity which is responsible for both the entrapment and the punishment. It seems to me plausible that we should place responsibility for the entrapment with the individual police officers and responsibility for punishment with the state.[14]

The last worry is that several of the views discussed here fail to account for the fact that there can exist instances of morally problematic entrapment outside the legal context as well.[15] Consider this case:

> *Fired:* Boss A dislikes his employee B and wants to fire him without a severance package. He decides to try to get B to commit a fireable offence which is sufficient to allow A to fire B without severance pay. He recruits another employee, C, to persuade B into breaking a company rule on the job. A is watching everything on CCTV and, just as B breaks the rule, goes to fire him.

Surely, A has entrapped B here. Moreover, it seems to me that, precisely because of this, it is also wrong for A to fire B without a severance package. But the previous views cannot so easily account for this. Contra the Legitimacy Account, for instance, A has not necessarily violated a precondition for his right to fire people. It does not seem that bosses, in general, need to be committed to 'the principles of justice' or to minimizing rule-breaking in order to have the right to fire employees for breaking the rules. Even bad and lazy bosses have a right to fire employees who break the rules. Furthermore, it is not necessarily the case that *the boss* would act incoherently by firing B. After all, it is not clear that A expresses that the offence is both fireable and not fireable. Indeed, he may have instructed C to express to B that the offence is fireable because he wants to have proof that B knew it was a fireable offence.

In sum, then, although existing views capture morally salient features, I believe these worries all suggest that we also need an account which focuses on the complaints of the targets, or victims, of entrapment and captures the sense that there is something *inherently* morally problematic about entrapment which makes punishing them wrong.

[13] This may even be the more realistic scenario in many jurisdictions.

[14] One reply is that as long as "agents of the state" are responsible for the entrapment, this is sufficient to claim that it is 'the state' which is responsible for it. I do not think this reply is convincing if the state really has done its best to outlaw entrapment, however. It may nevertheless be true that the state has a special responsibility for correcting the wrongs perpetrated by agents of the state, which is why it ought to not punish the entrapped offenders. But the views of entrapment outlined above would still not account for this. For it is not true that the *state* has violated its commitment to crime prevention, or *the state* which would be acting *incoherently*, for instance.

[15] Plausibly, some versions of the Standing Account avoid this worry because lack of standing makes blame inappropriate in interpersonal contexts as well.

The Concept of Entrapment

While the motivation for the Manipulation Account comes from concerns with existing accounts, the inspiration for the account comes from the *definition* of entrapment itself. It is often said that entrapment consists of (roughly) three parts:

(i) The police incite a target to commit a crime.
(ii) The target would not have committed the crime absent the incitement.
(iii) The police incite the target with the intention of having them arrested and punished.

The Manipulation Account I develop below holds that it is wrong to punish entrapped offenders because of (iii). This is the 'intentional aspect' of entrapment. Of course, the natural next question is what it means to intend for something. Here, I understand intentions along the lines of Yaffe's (2010) account, which in turn is inspired by Bratman's (1987) account. Put simply, intentions are practical mental states which 'play a role in deliberation and in the motivation and guidance of action' (Yaffe 2010, p. 53). More precisely, to intend that something *p* occur comes with rational and practical commitments: e.g., to pursue courses of action one believes are necessary to achieve *p* and not to deliberate about courses of action that one believes are incompatible with *p*. Moreover, when an intention that *p* occur plays its proper causal role, it is also what motivates one to pursue the actions in question. In the context of entrapment, we may say that the intentional condition (iii) is satisfied when the police are committed to realising the outcome in which the target is punished and that this commitment is what explains their other commitments and actions: trying to make the target commit the crime, intending to arrest them once they commit the crime, and so on.

It is surprising that, to my knowledge, no one has tried to solve the puzzle of entrapment by focusing on this 'intentional aspect' of entrapment. It is surprising because this purpose- or intention-element is included in most definitions of entrapment. For example, Ho says that it is an essential feature of entrapment that 'what motivates the operation from the start is the desire to have the person convicted and punished' (2011, p. 74). Duff et al. claim that entrapment involves inciting someone 'for the purpose of arresting and prosecuting him' (2007, p. 242). Hill, McLeod and Tanyi argue that it is necessary that the entrapping agent 'intends to be enabled, or intends that a third party should be enabled, to prosecute or to expose the target for having committed the act' (2018, p. 550). Similarly, Gerald Dworkin says that 'the central moral concern' regarding entrapment is that it manufactures crime 'in order that offenders be prosecuted and punished' (1985, p. 24).

Not everyone agrees, though. Stitt and James (1974) claim that there is entrapment as long as the police induce someone to commit a crime that they otherwise would not have committed.[16] But I think such conceptions of entrapment are mistaken. Consider, for instance, the following case inspired by Tadros (2005, pp. 318–319):

[16] See also Kim (2019, pp. 79–80) who thinks there can be negligent and reckless entrapment as well.

Failed Plan: Dirk, an undercover police officer, wants to arrest Charlie but has no criminal evidence against him. He approaches Harry—who dislikes Charlie—and encourages him to start an illegal fight with Charlie. His plan is to arrest Charlie, and only Charlie, once the fight breaks out. But before he has a chance to do so another police officer arrests both Harry and Charlie.

Although Dirk satisfies conditions (i) and (ii) with respect to Harry—i.e., he persuaded Harry to commit a crime he otherwise would not have committed—and Harry ends up being arrested, it does not seem to me that Dirk *entrapped* him. He did not lay out a 'trap' *for Harry*. By contrast, if we imagine that Dirk encouraged Harry to start an illegal fight for the sake of getting *Harry arrested*, then the case is instantly recognizable as a case of entrapment. As such, we have support for the view that the intentional element is necessary for entrapment. To echo Tadros's comments on a similar case, Dirk does not entrap Harry precisely because he does act '*in order to prosecute Harry*' (2005, p. 319).

Since the intentional element seems essential to the definition of entrapment, I believe this also gives us good reason to think that it is an essential part of the *explanation* for why punishing targets of entrapment is wrong. That is the inspiration for the Manipulation Account.

The Manipulation Account

Here is a simple account of entrapment which emerges naturally from what has been said so far. First, entrapment is entrapment partly in virtue of involving a plan *aimed* at the punishment of the target. Second, this is an *evil* or *wrongful* plan since the target is not a criminal to begin with. Third, it is morally bad if evil or wrongful plans are successful. Fourth, it is therefore bad if entrapped offenders are punished. The third premise is intuitively attractive but too vague.[17] For instance, it is not clear why it would be bad in itself if evil plans succeed. Moreover, this view would not account for the fact that targets of entrapment are *wronged* if they are punished. It is not just a bad outcome, or worse for the world—the targets in question have a reasonable moral complaint against being punished.

What I call the Manipulation Account is a more sophisticated version of this simple account. According to it, the intentional aspect of entrapment—the 'evil plan' aspect—makes entrapment manipulative in a way that also provides a manipulation-based reason not to punish entrapped offenders. In the rest of this section, I develop this claim in more detail.

To begin, let me highlight that understanding entrapment as involving manipulation is not novel. Hughes and Kim refer to it in rejecting the claim that targets of entrapment can be fully responsible for their crimes. Hughes talks about

[17] Both Parr (2016) and Duus-Otterström (2017) have claimed that it is bad if immoral plans succeed, but it is not really clear why they think this is bad.

manipulation-as-pressure which 'undermine the autonomy of those subject to them' (Hughes and Kim, 2004, p. 58). Kim similarly says that to manipulate someone '*is* to reduce their autonomy' (2019, p. 75). Both authors claim, then, that the manipulation involved in entrapment diminishes the target's culpability. As explained earlier, I do not think this kind of culpability-affecting manipulation is *essential* to entrapment.

Instead, I believe that entrapment always involves what I call *manipulation-by-hidden-intentions* which has previously been discussed in a more general context by Gorin (2014). This is a type of manipulation-by-deception. Deceptive manipulation often involves deception about things external to the manipulator, like when A manipulates B into eating dirt by deceiving him about its taste. But there are also instances of manipulation in which the manipulator deceives the manipulee only about his own intentions and motives. Consider:

> *Job Offer*: Linda has received a job offer at a great department, D1. Her colleague, John, encourages her to accept it, citing the genuinely good reasons for working at D1. But John suspects that another letter, from Linda's dream department D2, is on its way. He wants Linda to commit to D1 before it arrives because he wants to see Linda fare badly, and secretly hopes he will get an offer from D2 instead.

Intuitively, John manipulates Linda. But John does not lie to, or mislead, Linda about the good reasons for working at D1. Perhaps he deceives Linda about the prospects of getting an offer from D2. If Linda has the belief that 'I won't receive an offer from D2', then John might deceive Linda by *allowing* her to continue to have that false belief.[18] But we can stipulate this away. Suppose John and Linda both know there is a chance that Linda will get an offer from D2, but that Linda is ultimately persuaded by John's arguments to accept the first offer. John's behaviour still seems manipulative.

The best explanation for this, I believe, is that John engages in *manipulation-by-hidden-intentions*. John is manipulative in persuading Linda because he hides his real intentions: for her to fare badly and for him to benefit. To elucidate this idea, Gorin argues that there is a Transparency Norm governing communication. This norm 'requires that an interactive partner not hide her intentions in interacting when these intentions are relevant to the intentions and interests of the person with whom she's interacting' (2014, p. 78). Transparency about intentions is, in other words, an expectation when we interact with others. Violating this norm is therefore an instance of manipulation. According to Raz, for instance, manipulation 'perverts the way [a] person reaches decisions, forms preferences or adopts goals' (1988, p. 378). If we *expect* transparency about intentions when we make decisions based on persuasion by others, then violating that expectation is a way of perverting the way we reach decisions. In a similar vein, Gorin relies on a plausible idea from Buss (2005, p. 226), that an important feature of (many instances of) manipulation is that it prevents the manipulee from governing themselves with an accurate understanding of

[18] See, e.g., Chisholm and Feehan (1977).

their situation. Hiding one's true intentions can therefore be manipulation because, by doing so and "playing on the expectations of manipulees [...] manipulators prevent manipulees from governing themselves with an accurate understanding of their situation" (Gorin 2014, p. 78)

Although Gorin's claims here are attractive, it is an underdeveloped account. To claim that we are required to disclose our intentions whenever they are 'relevant' to the other person's interests and intentions is false. Sometimes, we are expected to do the opposite. Consider:

> *Proposal:* Jack and Jill have been a couple for years and Jack intends to propose to Jill today. He wants to propose at the spot they first met, so he comes up with a fake reason for why he needs Jill to meet him there later today. He then proposes, and they get married soon after.

In finding an excuse to get Jill to go to the spot, Jack is hiding his real intentions. Moreover, his intention—to propose—is also relevant to Jill's intentions and interests. But clearly, he has not *manipulated* Jill.[19] Indeed, social norms here seem at odds with Gorin's Transparency Norm. Social norms seem to demand that Jack actually hide his intentions—if not, it would ruin a great and romantic proposal.[20]

Spelling out the precise conditions for when the Transparency Norm holds is difficult, and beyond the scope of this paper. But it seems to me plausible to claim that the Norm at least applies when our intentions are relevant to the other person's interests *in a negative way*. That is, when the intentions that are hidden are aimed at something which is bad for the other agent. If one has been pushed towards making a decision by someone who secretly wanted something bad to happen to one, then one may rightly feel wronged and manipulated. Absent clear defeaters, it is natural to assume and expect that people at least do not have *bad intentions* for you when they encourage you to make a choice.

Now,what is important to highlight here is that this form of manipulation is wrongful also in cases in which it does not affect the culpability of the manipulee's actions. Imagine, for instance, that it is culpable for Linda to decide to take the job at D1—perhaps because it entails that she must abandon someone who is dependent on her staying. The culpability of that decision is not reduced by John's hiding his true intentions. It may make John extra culpable—since he is trying to get Linda to do something morally wrong—but it does not *reduce* the culpability of Linda's choice. Still, it is an instance of wrongful manipulation.

This is because, as Raz explains, the wrongness of manipulation (and coercion) 'transcends the severity of the actual consequences of these actions' (1988, p. 379). More precisely, Raz argues for an independence condition for autonomy:

[19] Some prefer a non-moral concept of manipulation which does not entail *pro tanto* wrongness. They might be happy to say that Jack manipulated Jill. But this merely pushes us to answer a different question: why is Jack's manipulation not wrongful while John's manipulation in *Job Offer* is? So, we are still forced to explain why keeping one's intentions hidden does not entail *wrongful* manipulation in all cases.

[20] I am grateful to Connor Kianpour for pressing me on this.

> [Independence] attests to the fact that autonomy is in part a social ideal. It designates one aspect of the proper relations between people. Coercion and manipulation *subject the will of one person to that of another*. That violates his *independence* and is inconsistent with his autonomy. (1988, p. 378; emphasis added)

There is something wrong with manipulation over and above the actual consequences (it may have) on our decisions and the deontic status of our decisions: the subjection of one will to that of another. This is precisely what seems to occur in *Job Offer*. John pushes Linda towards a choice for reasons that are hidden from her, in violation of her expectations. He pushes her towards a particular decision but prevents her from making that decision with an accurate understanding of her situation. In doing so, John subjects Linda's will to his own. Her will becomes a pawn in his game—she is used as a mere means in John's plan. It seems plausible that we have a general independence-based interest in being free from this kind of treatment—over and above the actual effects that manipulation may have on our decision-making and the deontic status of our decisions.

According to the Manipulation Account, entrapment always involves this kind of manipulation-by-hidden-intentions and is morally wrong for that reason. Consider the following case which is inspired by a genuine entrapment case which took place during the prohibition era in the United States:[21]

> *Sorrells*: A police officer, P, is introduced to S at S's home one night and plans to have him arrested. S is a World War I veteran. P tells S that he is a World War I veteran too, and they share stories. He also tells S that he is a police officer who is fed up with his work and could use a drink. P plays on their shared war experience try to persuade S to buy him a drink. S finally gives in and procures him a gallon of whisky. S is then arrested by P.

Clearly, P entrapped S, and this seems true regardless of the truthfulness of P's utterances. For instance, suppose that P is in fact a veteran, that he is genuinely fed up with his job and that he really does desire a drink. In that case, P does not deceive S about his identity as a police officer, about his war experiences or about his desire for alcohol. Still, he seems guilty of entrapping S.

This is explained by the presence of *manipulation-by-hidden-intentions*. In both versions of *Sorrells*, and in *Drugs*, the officer is hiding his true intention in encouraging the target to commit a crime. The true intention is, of course, that the criminal commit the crime and be arrested and punished. This intention is also hidden from the targets in all cases—if not, the entrapment would not be successful. So, although there may be many different kinds of problematic behaviour in different entrapment cases, what seems to unite them is that they all involve a particular kind of manipulation. This is the upshot of the previous section's argument that entrapment always involves a (hidden) intention for punishment and this section's argument that hiding one's true (negative) intentions is a form of manipulation.

[21] See Sorrells v. United States, 287 U.S. 435 (1932).

The Manipulation Account thus provides a clear argument for why it is wrong to entrap people: it involves manipulating people. But this alone does not show that it is wrong to subsequently *punish* targets of entrapment. That would follow if the following premise was true: if a police officer causes someone to commit a crime by acting wrongfully, then it is wrong to punish the offender for the crime. It would then be wrong to punish offenders for any crime created through entrapment (i.e., manipulation). To see that this premise is false, however, consider:

> *Red Light*: An undercover police officer, P, wrongfully runs through a red light. Seeing this, D decides to run through the red light as well.

Although D's crime resulted from P's wrongdoing, it does not seem impermissible to punish D. The premise above is false. So, the mere fact that entrapment involves something wrong (i.e., manipulation) does not entail that subsequent *punishment* is wrong.

The difference between *Red Light* and entrapment cases, however, is that the former case does not involve the particular kind of wrong that entrapment involves: manipulation-by-hidden-intentions. The presence of this kind of wrong explains why punishing the target in the entrapment case is wrongful. In short, entrapment, by being an instance of manipulation which intends for punishment, aims at punishment and therefore morally taints the morality of punishment. For the act of punishing the target is not simply the act of punishing a culpable criminal, it is also the act of *completing* or *fulfilling* the wrongful manipulation. And that is wrong.

One might object that the wrongful act of 'manipulating with a hidden intention' is finished as soon as the target is successfully manipulated into committing the crime regardless of whether punishment is subsequently imposed. If so, one may wonder how the imposition of punishment could be wrong for reasons related to the wrongfulness of the manipulation.[22]

The answer is that wrongs which are wrongs in virtue of intending for certain outcomes are aggravated by the realization of those outcomes. Only when those wrong are followed by the realization of those outcomes, are they as grave as they can be. This morally intimate connection between an act and certain consequences can be seen most clearly in cases in which the realization of an outcome partially *constitutes* the wrong in question. Consider the following case inspired by a case from Helen Frowe and Jonathan Parry (2019, p. 125):

> *Revenge*: After their break-up, Andreas decides to share nude photos of Betty, without her consent, on a 'revenge porn' website.[23] Carl and others later view the photos on the website.

As Frowe and Parry (2019, p. 126) explain, Andreas's wrongful act of *sharing the photos* depends on Carl and/or others subsequently looking at them. The outcome in which others look at the photos is not simply a causal consequence of Andreas's wrongdoing but partially *constitutive* of it. Moreover, the extent to which

[22] I am grateful to one of the judges of the *Res Publica* Postgraduate Essay Prize for pressing me on this.

[23] Revenge porn is '[s]exually explicit images or videos of an individual, published online without their consent and with the intent to cause them distress' (Chandler and Munday 2016).

this outcome is realized seems to make the wrong suffered by Betty graver—i.e., the more people look at the photos, the more seriously wrong Andreas's initial action is. This fact also helps explain why it is wrong for Carl and others to look at the photos: in doing so, they enable, and become complicit in, Andreas's wronging of Betty.

In my view, the realization of certain consequences can likewise aggravate wrongs which are wrongs partially in virtue of *intending* that those consequences are realized. To illustrate, consider the act of 'manipulative harming' someone, which is harming someone in a way that involves *using* them as a mere means to some end.[24] Manipulative harming is a particularly grave form of wronging, compared to, say, harming someone as a foreseen, but unintended, side-effect of some other action. On one popular view, it seems particularly wrong because it involves *treating*, or *using*, someone as a mere means. The concept of treating or using someone as a mere means to an end, moreover, requires an *intention* to use the harming of the victim or the harmful action as a way of achieving some goal.[25] So, the wrong of manipulative harming someone seems to consist in (i) the harm the victim suffers and (ii) the intention to harm them as a means to some end. This is not surprising. In general, to *use* something, like a tool, requires that one intends for it to play some role in fulfilling some end or reaching some goal. Now consider:

> *Enchanted Treasure*: An enchanted treasure requires a sacrifice of large amounts of human blood to be opened. Abby wants to get the treasure inside. At time *t1,* she kidnaps Bob and, against his will, draws a lot of his blood to use for the sacrifice, thereby making it possible for her to secure the treasure at *t2*.

Abby's *manipulatively harming* Bob is partially constituted by her intending the harmful act as a means to get the treasure. The gravity of that wronging can depend on consequences that lie in the future of Abby's initial actions here. For instance, it can depend on the extent to which harm-factor (i) is realized: the gravity of the wrongfully drawing of Bob's blood seems worse the more serious side-effects Bob develops over time because of it. But it can also depend on the extent to which the intention-factor (ii) is realized. Suppose, for instance, that we have a chance to intervene between *t1* and *t2*. That is, we cannot prevent the kidnapping and drawing of Bob's blood, nor the forming of the intention to use him, but we can prevent Abby from getting her hands on the treasure. It seems to me that there is a moral reason to intervene precisely because this will prevent Bob from being *successfully used* by Abby, which would be worse than being unsuccessfully used. In other words, the success of the plan would aggravate the wronging suffered by Bob, and that is why we should intervene.

[24] See, e.g., Tadros (2011, pp. 243–247, 2015).

[25] Kerstein (2013, p. 58) also emphasises that, to use someone as a mere means, one needs to intend for one's effect on them (e.g., harm) to contribute to reaching an end.

Further support for this comes from imagining a third party, Cam, who cannot do anything to stop Abby's actions but who will play some role in realizing the end. For instance, suppose that Cam will be responsible for bringing the treasure over to Abby after the sacrifice of Bob's blood. Intuitively, if Bob knew this and he could avoid being kidnapped by Abby by imposing some significant harm on Cam, then he would be permitted to do so. That is, he could impose some significant defensive harm on Cam. The most plausible explanation for this, I think, is that Cam's helping realize Abby's intended aim would aggravate the wronging of Bob and therefore make Cam complicit in Abby's wronging of Bob.[26, 27]

It is in this same way that, in the entrapment cases, the wrong of manipulation-by-hidden-intentions is aggravated by the extent to which the intended aim—punishment—is realized. Of course, this is not to say that that there is no wrong if the outcome is not realized. In *Enchanted Treasure*, we can blame Abby for a serious wrong (harming Bob and treating him as a means) even if she does not get the treasure in the end. Likewise, in entrapment cases we can blame the state for wrongful manipulation even if there is no punishment in the end. Still, in both these cases, the *graveness* of the wrong (manipulative harm or manipulation-by-hidden-intention) depends on the realization of certain outcomes. According to the Manipulation Account, then, Carlon is in many ways right when he says that when we refrain from punishing entrapped offenders we seek 'the prevention of a wrong's fulfilment' (2007, p. 1116).

This account, then, explains why it is wrong to punish entrapped offenders. If realizing some consequence, by performing some action, will aggravate the wronging of someone, then there is a duty to refrain from realizing the consequence. That is why the state has a duty to refrain from punishment in the entrapment context. This duty holds for anyone whose actions would realize the aggravating consequences, but it is plausibly strongest for those who are responsible for the primary wrongful act as well. As such, we can account for the sense that punishing entrapped agents is particularly problematic when it is one and the same state involved, but that it is also morally problematic to punish the entrapped offender in *Treaty*. Moreover,

[26] An alternative explanation of these intuitions may be that Abby should not be permitted to *benefit* from her wrongdoing, which is why we should intervene, and Cam may be liable in virtue of helping Abby benefit from her wrongdoing. But this fact alone cannot explain why there seems to be a reason to intervene also if Abby's plan was to benefit an innocent, unsuspecting person instead. I do not think the act of passively and involuntarily benefitting from an unjust act is itself wrong. So, the reason to intervene cannot be explained by the fact that it would prevent a separate wrong.

[27] One objection goes as follows. Suppose someone lethally and wrongfully pushes a man off a bridge intending for his body to stop a trolley from killing five other people. But suppose that we can intervene after the man has died. We can ensure that the trolley still kills the five. If I am correct above, it seems that there is a *pro tanto* duty to intervene and ensure that the five will die. In doing sp, after all, we would be preventing the realization of the intended outcome of the wrongful, manipulative killing of the first person. But that seems counterintuitive. My response is that there is no all-things-considered duty to intervene. The fact that all the harm is already suffered by the man and that intervening would only be bad for five innocent people suggest that we should not intervene. Still, I am sympathetic to the idea that there is *something*, grounded in a concern for the man, which (defeasibly) pushes against the realization of this outcome.

we can account for the fact that there is an *intimate connection* between the reasons not to entrap and the reason not to punish those who are entrapped, and that punishing an entrapped offender is wrong in part because it *wrongs* them. The former follows from the fact that, according to the Manipulation Account, the duty not to entrap and the duty not to punish are both ultimately grounded in the duty not to manipulate people. The second follows because this is a duty that we *owe* to people. On the Manipulation Account, then, it is not simply the case that D is culpable but the state has compromised the standing, authority or integrity required to punish them. Instead, although D cannot complain about the punishment *qua* being a culpable offender, he can complain about the punishment *qua* it being the fulfilment of a wronging—manipulation—that he was not originally liable to or deserving of. Lastly, this view is also better able to account for how entrapment is problematic in general, also outside of the legal context. For example, the moral problem in *Fired* is that by actually firing B, A will fulfil the wrongful manipulation he began subjecting B to when he sent C to persuade him into committing a fireable offence.

Objections

Although the Manipulation Account is a plausible theory of what is wrong about punishing entrapped offenders, it has some untraditional implications. In this last section, I outline three untraditional implications that may be considered objectionable and explain why they should not be considered objections after all.

Private Entrapment

It is often said that any plausible theory of entrapment must avoid the Problem of Private Entrapment.[28] No theory of entrapment should entail that it is wrong to punish those who have been encouraged to commit crimes by other *private citizens*. For instance, it is not wrong to punish someone simply because he was encouraged to commit a crime by a friend.

Initially, it does not seem that the Manipulation Account has a problem here. Private citizens who persuade others to commit crimes will not often *aim* for the punishment of the other person. Most likely, they will want them and themselves to walk free or, at worst, be indifferent about what happens to the other person. But we can conceive of more problematic cases, such as the following:

> *Envy:* James loves Amy but Amy loves Dylan. James devises a plan to get Dylan out of the picture by having him sent to jail. He knows that James is looking at serious prison time if he is arrested now because, although he is reformed, Dylan has been punished for several crimes in the past. James calls the police to report a crime and then begins persuading Dylan to commit a crime. As Dylan commits the crime, the police arrive and arrest him.

[28] See, e.g., Carlon (2007) and Yaffe (2005).

This is a case of *proper* private entrapment because a private citizen incites a crime for the purpose of having the other person arrested and punished. Since there is manipulation aimed at punishment, moreover, the Manipulation Account entails that Dylan's punishment would be morally tainted. The account therefore does not completely avoid the Problem of Private Entrapment.

However, I do not think that this case is fatal for the Manipulation Account. To see this, we can start by distinguishing two kinds of scenarios: one in which Dylan is manipulated into committing a minor, victimless crime and one in which he is manipulated into committing a serious crime with innocent victims. In the latter case, I think it is compatible with the Manipulation Account to claim that it is morally permissible to punish him. Recall, it is an account only of the *pro tanto* wrongness of punishing entrapped offenders and, so, is compatible with the existence of reasons to punish which can outweigh the entrapment-based reason not to punish. As an example, imagine that the crime in question harmed or significantly disrespected an innocent third party. In that case, the state may have a reason, owed to the victim of the crime, to hold Dylan to account by punishing him.

One argument for this begins with the plausible idea that punishment is partially about communication and expression and that it gets part of its value and justification from that fact.[29] It publicly condemns the offender and communicates, to both offender and others, the community's disapproval of the criminal behaviour. In doing so, however, it also communicates and expresses something about *the victim* of the crime: that they have moral status and worth and that what happened to them was wrong and worthy of condemnation because of how it affected the victim. Arguably, in standing up for the victim and reaffirming their status and worth by punishing the criminal, the state does something that is morally valuable and important for the victim.[30] There is therefore a powerful reason to punish the privately entrapped Dylan in this case, and this reason can plausibly outweigh the entrapment-based reason not to punish the offender.

This argument does not entail that it will be equally easy for the state to overcome the entrapment-based reason not to punish when the state itself is responsible for the entrapment. According to the Manipulation Account, the duty not to punish an entrapped offender is a duty to mitigate, or avoid worsening, the gravity of the initial wrong. But, as suggested earlier, these kinds of duties—such as the duty to undo wrongdoing, if possible, or to compensate for wrongdoing—are in general more stringent when it is one's own initial duty that has been breached than when it is another person's initial duty that has been breached. Take Satz's (2012, p. 137) example: A steals B's bike one day, and after they both die their sons learn what happened. Plausibly, A's son has a compensatory duty to give the bike to B's son now. He should try to mitigate his father's initial wrongdoing. Still, his duty to

[29] For more on the expressive and communicative function and role of punishment, see, e.g., Feinberg (1965) and Duff (2001).

[30] See, e.g., Statman (2008) for more on the value and importance of communicating the moral status of victims to wrongdoers, third parties and victims themselves. He argues that the value of this can justify so-called futile defensive force, for instance. See also Alm (2019) for an argument in favour of there being reasons, owed to victims, to punish offenders which is also focused on the communicative aspect of punishment.

return the bike is weaker than A's duty to return the bike. For instance, it would be easier for the son to justify not doing so on the basis of costs to himself than it would be for A, precisely because the duty is not a duty to compensate for *his own* breach.

The same is true with respect to manipulation-by-hidden-intentions and entrapment. The duty not to punish when one is responsible for the initial entrapment is stronger than the duty not to punish when another entity or agent is responsible for it. That is why one and the same moral consideration—e.g., the expressive value of standing up for the victim's status and worth—can outweigh the reason not to punish an entrapped offender in some cases but not in others. More precisely, it is why that moral consideration can more easily outweigh the entrapment-based reason not to punish in private entrapment cases than it can in state entrapment cases.[31] So, this response to the *Envy* case should not be read as one which also weakens the duty not to punish in state entrapment cases. The Manipulation Account is compatible with holding that the duty not to punish is much stronger in those cases.

In versions of *Envy* in which Dylan is manipulated into committing a lesser crime, in which there are no innocent parties harmed or disrespected, the Manipulation Account suggests that the state should not punish him. So, it does allow for private entrapment to make punishment wrong. But I think that we, in fact, have reason to welcome this implication of the account.[32] First, we should not want it to be (easily) possible for wrongdoers to subvert or co-opt our criminal justice system to further their evil plans. Yet this is precisely what we would allow if we insist that Dylan should be punished: James would have successfully co-opted our justice system to further his unjust plan to get Dylan out of a love triangle. To put the point even stronger, there is a risk that, if we punish Dylan, then we—or the state—become *complicit* in James's wrongful plan because we—or the state—would play an active and important role in his plan. Again, we have reason to want to avoid this and, so, have reason to welcome the implication of the Manipulation Account.

However, some may worry that this implication of the Manipulation Account is objectionable for a different reason. It is not that it allows for the possibility that punishment is impermissible due to private entrapment per se, but rather that in doing so it commits us to a radical view about the moral importance of mental states. Consider *Envy* again and compare it to a case in which all facts about the manipulation and the crime are identical but we remove the *entrapping* element. For instance, imagine that John uses the same methods as James to get someone else, Dan, to commit a crime but that he has no intention for Dan to be punished. He does

[31] Of course, in state entrapment cases the state also has an extra strong reason to ensure that the status and worth of the victim is communicated and expressed insofar as the state is co-responsible for the wronging the victim has suffered. Yet this need not be an extra reason to punish the entrapped offender. Indeed, it may be more of a reason for the state to hold itself accountable in some way and likely to compensate the victim as well.

[32] I am not the only one who is not too worried about this possibility. Ho (2011, p. 92), for instance, thinks that it is wrong for the state to punish in *private prosecution* cases in which a private individual has entrapped the offender. Furthermore, Dein and Collier (2014) and Stark (2018) discuss, and offer some support for, actual cases in which a stay of prosecution has been granted on the basis of private entrapment.

it only because he thinks he and Dan stand to benefit from the crime. According to the Manipulation Account the state should punish Dan but not Dylan since only one of them was the target of private entrapment. More precisely, the only reason that the state should treat the two radically different is that James *intended* for Dylan to be punished while John did not *intend* for Dan to be punished. That may strike some as counterintuitive.

I do not think this worry provides reason to reject the Manipulation Account. First, if the worry is that we should treat Dan and Dylan radically different based on differences in other people's mental states, then this is a general worry for *all* accounts of entrapment which make the *intention* or *plan* to arrest and punish the target part of the definition of entrapment. As we saw, the majority of accounts of entrapment make the intentional element an essential part of the concept. So, even within the domain of state entrapment, most accounts are committed to treating people differently based on differences in the mental states of the agents of the state. For those differences determine whether some action is entrapment or not. Thus understood, the objection casts a much wider net and does not single out the Manipulation Account as problematic.[33]

Furthermore, I do not think the suggestions that mental states can significantly affect the moral landscape is so counterintuitive. Take, for instance, the issue of complicity, and consider Ava who is standing outside a bank while a robbery is taking place inside. From her position, she can easily see if and when the police arrive and alert the robbers inside, but no police show up before the robbers are all gone. Whether Ava was complicit in the robbery is morally, and legally, significant as it determines whether she should be punished and, if so, how much. But whether she was complicit seems to depend, to some significant degree, on her mental states—in particular, her intentions and plans. Although she did not make a difference to the robbery—since the police did not show up—we may plausibly call her complicit if her *intention* was to alert the robbers inside if she saw the police.[34] So, the fact that differences in mental states can have significant moral consequences should not be surprising.

Of course, one difference here is that it is Ava's mental states which matter for what it is right or wrong to do with respect to Ava herself. In the previous cases, the suggestion is that the mental states of James and John matter for what it is right or wrong to do with respect to someone else: i.e., Dylan and Dan. That may be more radical. But consider Ava again. On some views, Ava's intention to help is insufficient for complicity. It is also essential that the *others* intend for Ava to fulfil the role that she intends to fulfil.[35] A stranger who merely intends to help some wrongdoers is not *complicit* in their crimes. On that view, then, what is right or wrong to do with respect to someone can depend on the intentions and mental states of other people.

[33] If anything, it presents an objection to all intention-based accounts of entrapment. I do not think it succeeds, but a full blown defence of the intention-based accounts of entrapment is beyond the scope of this paper.

[34] For more on the significance of intentions to determining complicity, see, e.g., Kutz (2007) and Bazargan (2013).

[35] For instance, Bazargan (2013, pp. 186–87) highlights the importance of *shared* intentions in complicity cases.

Alternatively, consider again the distinction between manipulative killing and killing as a side-effect. Most moderate deontologists hold, for example, that it can be permissible to kill someone as a side-effect of saving some number of lives—say, five—yet impermissible to kill someone as a means of saving the same number of lives. So, it may be permissible to divert a trolley away from five people in the knowledge that it will kill one other person on the other track, but impermissible to push someone onto the track for the sake of using their body to stop the trolley from killing the five. Suppose we accept this, as I think we should. As said earlier, whether you are using someone as a means depends on whether you intend for or plan to use them for the sake of achieving something. Suppose, then, that A is impermissibly trying to manipulatively kill V1 as a means of saving five lives while B is about to kill V2 as a side-effect of saving five other lives. Now, suppose we can save only V1 or V2 but not both. It seems to me quite plausible to think we should save V1 over V2. The former's death, after all, is contrary to the demands of morality while the latter's death is compatible (indeed, perhaps favoured) by the demands of morality. Thus, it should not be surprising if morality prefers that V1's death is prevented. Yet, this conclusion about whom to save will, ultimately, be due to differences in the intentions and mental states of A and B. Again, then, we should not find it so surprising that differences in mental states, in Dan's and Dylan's cases, can make a significant moral difference.[36]

To sum up, then, I do not think the private entrapment concern is a reason to reject the Manipulation account. In cases of serious crimes, the Account has the resources to side with the more traditional accounts of entrapment in holding that the offender should ultimately be punished. Moreover, I have argued that the fact that the state should not punish in cases of minor crimes should not be seen as an objection to the Account.[37]

Mere Opportunities

A different objection holds that the Manipulation Account over-generalizes and wrongly entails that it is wrong to punish targets of every kind of *proactive* policing. Consider:

> *Pickpocket*: A certain local area has seen a drastic increase in pickpocketing. The police send out an undercover officer, P, with a wallet visibly sticking out

[36] Admittedly, the manipulative killing example is not fool-proof support since there exists alternative ways, not similarly sensitive to intentions, of spelling out the particular wrongness of manipulative harming. For a great overview and discussion, see Ramakrishnan (2016).

[37] I am grateful to an anonymous reviewer of this journal for pressing me on my original response to the issue of private entrapment. They also wondered what the evidence for proving private entrapment would have to be given that it is a matter of people's intentions. I confess it is difficult to think of evidential standards. But there seem to be cases in which we can quite plausibly deduce that there was private entrapment. See, e.g., Dein and Collier (2014) and Stark (2018). Moreover, as a rule of thumb, the following seems quite plausible. If A (i) induces B to commit a crime, then (ii) immediately turns around and alerts the police, and (iii) we cannot find evidence of any other reason why A would behave in this way, then we should seriously suspect private entrapment.

of his back pocket in an effort to lure out the pickpockets and arrest them. D grabs the wallet and tries to run away, but he is quickly apprehended.

Many people hold that this is permissible proactive policing, not wrongful entrapment. Existing legal doctrines concerning entrapment are designed to account for this. According to the Subjective Test for entrapment, proactive policing is not entrapment if the target was *pre-disposed* to commit the crime. According to the Objective Test, proactive policing is entrapment only if the tactics used are so serious that they would have caused most reasonable, law-abiding citizens to commit a crime as well. Both tests are designed to avoid the conclusion that it is wrong to punish those who grab 'mere opportunities' to commit crimes that are presented by the police.

The Manipulation Account seems incompatible with both tests. If the police intended that someone seize the opportunity and be punished in *Pickpocket*, then there is manipulation aimed at punishment and the punishment is morally tainted. However, the fact that the Manipulation Account is at odds with the two tests is not problematic, I believe, because we should reject them.[38] I agree with Stitt and James who claim that '[n]o one should be offered an opportunity to commit a crime unless there's probable evidence that he's engaged in ongoing criminal activity' (1974, p. 130).

That claim points towards a different test, which we may call the Prevention Test. According to this, proactive policing is entrapment (and wrong) whenever a criminal opportunity is presented (with the relevant hidden intentions) *unless* there is a preventive justification for doing so. The preventive justification in question is present when there is a significant likelihood that the target will commit a crime (of similar, or greater, severity) at some other time when the police will not be able to arrest him (at least not so easily). Importantly, this exception to the rule—i.e., that it is not wrongful entrapment if the target is likely engaged in, or about to be engaged in, criminal activity—is not *ad hoc*. Underpinning it is the idea that one can become *liable* to *prima facie* wrongful treatment if doing so is necessary to prevent one from doing something wrong. The reason why it can be permissible to present criminal opportunities to those suspected of being criminals is that they are liable to this kind of manipulation.[39]

The Preventive Test is compatible with the Manipulation Account because manipulation is not wrongful when targeted at liable agents. It is therefore also permissible to punish the targets of the manipulation in those cases. For that reason, the

[38] This is not a novel position. Stitt and James (1974), Dworkin (1985), Howard (2016) and Lippke (2017) are all sceptical of both tests.

[39] I borrow here from Nathan's (2017) argument that people engaged in criminal activity can be liable to sting operations and therefore not wronged by the deception and manipulation often involved. Of course, there are complications here concerning the standard of proof. Liability claims are least contentious when we are certain that someone is engaged in criminal activities. In sting operations and entrapment scenarios, however, one problem is that we often cannot know with certainty whether the target is engaged in, or about to be engaged in, criminal activities. There is therefore a risk that we will manipulate an innocent person. Plausibly, then, the standard of proof should be quite high.

Manipulation Account is consistent with thinking that it is permissible to punish D in *Pickpocket*. If the police put out the bait somewhere with a significant pickpocketing problem, their actions can satisfy the Preventive Test. It is only impermissible to punish the target if the police had no preventive justification for putting out the bait at that location. This, I think, is the intuitively correct view as well.

Virtue Testing

The Manipulation Account gets a lot of entrapment cases right. Provided there is an intention to have the target punished, there is a manipulation-based reason to refrain from punishment. This is true whether the entrapment is done for general deterrence reasons, sadistic reasons (e.g., a police officer who simply wants to see a person suffer punishment) or prudential reasons (e.g., a police officer who hopes he will get a promotion by sending more people to prison). But consider:

> *Virtue Testing*: An undercover police officer, P, encourages D to commit a crime. He hopes that D will *not* commit the crime, but he intends to arrest and have D punished *if* D commits the crime.

The virtue testing police officer does not intend *that* D commit a crime *and* be punished. He only conditionally intends for punishment and hopes that the condition will not be satisfied.

The lack of a non-conditional intention *that* D be punished suggests that P's action does not *aim* at punishment. Consequently, the punishment of D will not be morally tainted according to the Manipulation Account. Some might argue that this is a problem for the account. Indeed, some believe that entrapment is objectionable precisely because it is a form of virtue testing.[40] They might therefore insist that it is wrong to punish in *Virtue Testing* as well.

There are two responses to this objection. The first is to reject the claim that it is wrong to punish in *Virtue Testing*. After all, the case does not seem to fit the most plausible definitions of entrapment which, recall, all include an intentional element. *Virtue Testing* is more like *Failed Plan* and *Red Light*. Moreover, the Manipulation Account can still explain why virtue testing *itself* is wrong. P's virtue testing is *manipulative* because he's hiding his real intentions from D (i.e., that he is encouraging D for the purpose of *testing* his virtue).

The second response is to accommodate the claim that it is wrong to punish in *Virtue Testing*. In contrast to *Failed Plan*, and *Red Light Virtue Testing* involves an intention that D be punished for his crime. It just happens to be a *conditional* intention. As is familiar from the criminal law context, conditional intentions should sometimes, but not always, be treated the same as unconditional intentions. Consider the person who enters someone else's house with the intention to steal *if* they find something valuable enough but who does not find it.[41] Although laws against

[40] See, e.g., Tunick (2011) and Dworkin (1985).

[41] Inspired by *Regina v Greenhoff* [1979] Crim LR 108, discussed in Campbell (1982).

burglary tend to require the *intent* to steal, surely this person qualifies as having tried to burglarize the house even though his intention to steal was conditional. But imagine now a person who hijacks a car with a another person in it with the intention of killing the passenger *if* the police are alerted and, get too close to them and the carjacker feels he has no other option. It seems much less obvious that he ought to be found guilty of carjacking *with* the intent to kill.

Again, inspired by Bratman, Yaffe (2004) offers a compelling way of thinking about these cases and determining when someone's conditional intention should be considered to fit the *mens rea* of a more serious crime. Recall, on this theory of intention, intentions come with various rational commitments that structure one's deliberations. For instance, intending to x rationally commits one to pursue actions that make x-ing possible, not to form intentions to do things incompatible with x-ing and so on. In simple terms, Yaffe's idea is the following. To figure out whether someone who conditionally intends to do x should be treated as someone who unconditionally intends to do x or someone who lacks an intention to do x, we should look at how his rational deliberations are structured. Are they more similar to the rational deliberations of the former or the latter?

A carjacker with an unconditional intention to kill the passenger will be guided by particular deliberations and commitments. He is likely to have ensured the gun is loaded, prepared for the passenger's possible escape attempts, thought about how to dispose of the body and so on. Someone who merely conditionally intends to do so may be quite different. In the extreme case, the person may only have briefly considered the circumstances in which he intends to kill in (e.g., the police catching up to them). However, they may be very similar to the carjacker with the unconditional intention to kill. For instance, they may believe that it is very likely that the police will catch up to them, have deliberated a lot about what to do in that event, and made preparations for that outcome (like making sure the gun is loaded, that the victim cannot easily escape, and so on). It seems plausible that we should find the latter carjacker (but not the former) guilty of carjacking with the intent to cause death given how similar their practical deliberations and commitments are to the carjacker with the unconditional intention to kill.

The same approach can be used to determine when cases like *Virtue Testing* ought to be treated as entrapment. Typically, I think, the and commitments of a virtue testing police officer will be similar to those of a standard entrapping police officer. For instance, they will both likely have deliberated about what to do if the target commits the crime and have taken steps to facilitate an arrest in that event, and they are both likely to be committed not to do things incompatible with the target being punished in the end. The fact that the virtue testing police officer also manipulates the target, and thus makes it more likely that the target will commit a crime and he will have to facilitate an arrest and so on, makes it appropriate to treat him as sufficiently similar to the standard entrapping police officer who unconditionally aims for the target to be punished.

Compare this to a conditional intention version of *Failed Attempt*. Recall, this is the case in which a police officer, Dirk, tries to persuade Harry to fight Charlie because he wants to arrest *Charlie*. I said Dirk did not entrap *Harry* because he did not act for the sake of getting Harry arrested and punished. But we can imagine a case in which he conditionally intends for this. Suppose that, by coincidence, Harry is white and Charlie is Black and that Dirk is worried he may be in trouble if someone sees him only arresting the Black person involved in the fight. So, he conditionally intends to also arrest Harry as well *if* he notices that someone is filming the arrest. Still, if we find that Dirk is guided in large part by deliberations about how to reduce the likelihood of this condition being fulfilled—e.g., that he plans to make the arrest when no one else is around—it does not seem that we should treat Dirk as similar to a standard entrapping police officer who would encourage Harry to fight for the sake of arresting and punishing him.

Ultimately, then, I do not think *Virtue Testing* is a counterexample to the Manipulation Account. First, we may reasonably think they punishment is not appropriate in this case. Second, even if it is objectionable, the Manipulation Account can likely be extended to cover that judgement as well if we adopt an approach to conditional intentions like Yaffe's.[42]

Conclusion

The puzzle of entrapment is generated by seemingly conflicting intuitions concerning the wrongness of entrapment and the wrongness of punishing entrapped offenders on one hand and the culpability of the offender on the other hand. After outlining some novel objections to many of the views in the existing literature, I developed a new solution to the puzzle grounded in what I called the Manipulation Account. A virtue of that account is that it takes the definition of entrapment seriously, and it is able to account for the sense in which it is inherently problematic to punish those who are victims of entrapment *because* they are victims of entrapment. To punish an entrapped offender is to fulfil or complete, and thereby aggravate, the wrongful manipulation that they are victims of. This is why there is a *pro tanto* reason not to impose punishment on them, even though they may be fully culpable.

[42] I am grateful to an anonymous judge for the *Res Publica* Postgraduate Essay Prize for helping me improve my response to the virtue testing objection.

References

Alm, D. 2019. Crime victims and the right to punishment. *Criminal Law and Philosophy* 13: 63–81.

Ashworth, A. 1999. What is wrong with entrapment? *Singapore Journal of Legal Studies* 40: 293–317.

Bazargan, S. 2013. Complicitous liability in war. *Philosophical Studies* 165 (1): 177–195.

Bratman, M. 1987. *Intention, plans and practical reason*. Cambridge, MA: Harvard University Press.

Buss, S. 2005. Valuing autonomy and respecting persons: Manipulation, seduction, and the basis of moral constraints. *Ethics* 115 (2): 195–235.

Campbell, K. 1982. Conditional intent. *Legal Studies* 2: 77–97.

Carlon, A. 2007. Entrapment, punishment, and the sadistic state. *Virginia Law Review* 93: 1081–1134.

Chandler, D., and R. Munday. 2016. *A dictionary of social media*. Oxford: Oxford University Press.

Chisholm, R., and T. Feehan. 1977. The intent to deceive. *Journal of Philosophy* 74 (3): 143–159.

Cohen, G.A. 2006. Casting the first stone: Who can, and who can't, condemn the terrorists? *Royal Institute of Philosophy Supplement* 58: 113–136.

Dein, J., and L.V. Collier. 2014. Non-state agent entrapment—The X factor. *Archbold Review* 9: 4–6.

Dillof, M. 2004. Unraveling Unlawful entrapment. *Journal of Criminal Law and Criminology* 94: 827–896.

Duff, R.A. 2001. *Punishment, communication and community*. Oxford: Oxford University Press.

Duff, A., L. Farmer, S. Marshall, and V. Tadros. 2007. *The trial on trial*, vol. 3. Cornwall: Hart Publishing.

Duus-Otterström, G. 2017. Benefiting from injustice and the common-source problem. *Ethical Theory and Moral Practice* 20: 1067–1081.

Dworkin, G. 1985. The serpent beguiled me and I did eat: Entrapment and the creation of crime. *Law and Philosophy* 4: 17–39.

Feinberg, J. 1965. The expressive function of punishment. *The Monist* 49: 397–423.

Frowe, H., and J. Parry. 2019. Wrongful observation. *Philosophy and Public Affairs* 47: 104–137.

Gardner, J. 2011. What is tort law for? Part 1. The place of corrective justice. *Law and Philosophy* 30: 1–50.

Gorin, M. 2014. Towards a theory of interpersonal manipulation. In *Manipulation: Theory and practice*, ed. C. Coons and M. Weber, 73–97. New York: Oxford University Press.

Hill, D.J., S.K. MacLeod, and A. Tanyi. 2018. The concept of entrapment. *Criminal Law and Philosophy* 12: 539–554.

Ho, H.L. 2011. State entrapment. *Legal Studies* 31: 71–95.

Howard, J. 2016. Moral subversion and structural entrapment. *The Journal of Political Philosophy* 24: 24–46.

Hughes, P.M. 2004. What is wrong with entrapment? *The Southern Journal of Philosophy* 42: 45–60.

Kerstein, S.J. 2013. *How to treat persons*. Oxford: Oxford University Press.

Kim, H. 2019. Entrapment, culpability, and legitimacy. *Law and Philosophy* 39: 67–91.

Kutz, C. 2007. Causeless complicity. *Criminal Law and Philosophy* 1: 289–305.

Lippke, R.L. 2017. A limited defense of what some will regard as entrapment. *Legal Theory* 23: 283–306.

Nathan, C. 2017. Liability to deception and manipulation: The ethics of undercover policing. *Journal of Applied Ethics* 34: 370–388.

Parr, T. 2016. The moral taintedness of benefiting from injustice. *Ethical Theory Moral Practice* 19: 985–997.

Ramakrishnan, K. 2016. Treating people as tools. *Philosophy and Public Affairs* 44 (2): 134–165.

Raz, J. 1988. *The morality of freedom*. Oxford: Oxford University Press.

Satz, D. 2012. Countering the wrongs of the past: The role of compensation. *Nomos* 51: 129–150.

Stark, F. 2018. Non-state entrapment. *Archbold Review* 10: 6–9.

Statman, D. 2008. On the success condition for legitimate self-defense. *Ethics* 118 (4): 659–686.

Stitt, G., and G. James. 1974. Entrapment and the entrapment defense: Dilemmas for a democratic society. *Law and Philosophy* 3: 111–131.

Tadros, V. 2005. *Criminal responsibility*. Oxford: Oxford University Press.

Tadros, V. 2011. *The ends of harm: The moral foundations of criminal law*. Oxford: Oxford University Press.

Tadros, V. 2015. Wrongful intentions without closeness. *Philosophy and Public Affairs* 43 (1): 52–74.

Tadros, V. 2020a. Secondary duties. In *Civil wrongs and justice in private law*, ed. J. Oberdiek and P.B. Miller, 185–207. Oxford: Oxford University Press.

Tadros, V. 2020b. Distributing responsibility. *Philosophy and Public Affairs* 48 (3): 223–261.
Tunick, M. 2011. Entrapment and retributive theory. In *Retributivism: Essays on theory and policy*, ed. M. White, 171–191. Oxford: Oxford University Press.
Wallace, R.J. 2010. Hypocrisy, moral address, and the equal standing of persons. *Philosophy and Public Affairs* 38 (4): 307–341.
Yaffe, G. 2004. Conditional intent and *Mens Rea*. *Legal Theory* 10: 273–310.
Yaffe, G. 2005. "The government beguiled me": The entrapment defense and the problem of private entrapment. *Journal of Ethics and Social Philosophy* 1: 1–50.
Yaffe, G. 2010. *Attempts: In the philosophy of action and the criminal law*. Oxford: Oxford University Press.