Entrapment and Manipulation[*]

Jonas Haeg

I. 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 reasons. On the one hand, most people believe it is wrong to entrap D and 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 satisfies 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 convinced by a *genuine* drug dealer. D appears just as culpable for the crime as this counterpart would be and, so, seems 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. This paper aims to resolve the apparent tension between these two intuitions 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. We have *pro tanto* duties to avoid completing wrongs such as manipulation. Hence, punishing entrapped offenders is *pro tanto* wrong.

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 where it is all-things-considered

[*] 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 Res Publica Postgraduate Essay Prize judges, as well as an anonymous reviewer of this journal.

permissible to punish an entrapped offender. For instance, it may be permissible to punish targets of entrapment who have been radicalised by undercover agents and pose a serious danger to society if not punished for their crime. Or, if the crime in question was 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 punishment. 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 because 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 entrapment and permissible proactive 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 proceeds as follows. Section II motivates the need for a new solution to the puzzle of entrapment by outlining some novel objections to existing solutions. Section III argues that entrapment always *aims* at punishment. Building on that, Section IV develops The Manipulation Account. Section V considers some potentially objectionable implications of the account and argues that they are not objectionable. Second VI concludes.

II. 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 entrapment constitutes a puzzle. The puzzle, recall, is generated by two seemingly incompatible intuitions elicited by cases like *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). Jeffrey 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 like *Drugs*. Paul M. Hughes (2004) and Hochan Kim (2019) are both sceptical of (2). They think that entrapment necessarily 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, so they are not liable to full punishment. Although I agree that entrapment *often* involves severe pressure, 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 the drug

[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 Section V.B.

dealer is genuine. It does not seem that the mere fact that this dealer persuaded D to commit a crime sufficiently undermines D's responsibility.[2]

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. It is these problems that motivate the need for an account like the Manipulation Account. To see this, let me briefly describe a few existing accounts.[3]

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.[4] Some think this framework helps explain why it is wrong for the state to punish entrapped offenders. Hock Lai Ho (2011), for instance, argues that the state lacks the standing to blame and condemn offenders for crimes the state has wrongfully instigated or caused – i.e., those 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.[5] 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, Andrew Carlon (2007) argues that a precondition for the "right of prosecution" is that a state embodies the principles of justice but claims that states which entrap 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 outlined by Antony Duff, Lindsay Farmer, Sandra Marshall, and Victor Tadros (2007).[6] 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, 236). More precisely, they argue that a problematic incoherence occurs when a state tries to punish someone they have entrapped. The reason is that, in entrapping the person, the state is – through its

[2] Indeed, it would be troubling if it did. People would be afforded (partial) excuses left and right if they could demonstrate that someone else persuaded them into committing a crime. Only those who motivate themselves to commit crimes would be liable to full punishment.

[3] For a great overview of even more accounts, and some of their problems, see Howard (2016).

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

[5] To clarify, Kim thinks 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 punishment is justified on grounds other than blame, such that a state could justifiably punish without having the standing to blame.

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

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. It is therefore wrong for states which entrap to punish because they would behave incoherently by doing so, which would undermine the integrity of the trial and criminal justice system.

These accounts and others capture different features that all seem relevant to explaining what is wrong with punishing entrapped offenders. But 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 give us correct verdicts about. Let me start with the second complaint. In looking at many of the accounts of entrapment, it is easy to feel that people have lost sight of, arguably, the most important person and the most important action: the target of the entrapment and the act of entrapment. Both things do little work in explaining why it is wrong for the state to punish someone, like D, who 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 ensuring the *integrity* of the state, and the fact that incoherent behaviour undermines it. This understanding is also reinforced by Andrew Ashworth's version of the view. 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). 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 due to the state's lack of authority. As such, 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).

This is also true of the Standing Account. In outlining the rationale for his version of this view, Ho explicitly asserts that 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 citizens to account for a certain crime (i.e., one that it has helped create). Again, the main focus in explaining why it is wrong to punish D is that the state has morally tainted 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, it is wrong to ignore, in the way they do, the *targets* of the entrapment. Intuitively, part of the explanation of why it is wrong to punish D is precisely that it *wrongs D* and, moreover, that it wrongs D because he was wrongfully *entrapped* into committing the crime in question. 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 do punish him. This cannot be so easily captured by accounts that only focus on the state's lack of authority, integrity or standing. A state can come to lack these things in other ways without giving the prospectively punished 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. Per the previous accounts, the state may now lack integrity or the authority and standing to punish X for his crime. But 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. 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 continuity between the reasons not to punish D and the reasons not to entrap him in the first place.[7] To see this, notice, for instance, that the severity of the wrongness of entrapment appears connected to the prospects 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. It may still be somewhat wrong to entrap D into breaking the law.[8] But it seems much less wrong than it would be if punishment were likely to follow. D's complaint against the state's entrapment seems much 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 mitigates the severity of the initial wrong. Again, the previously mentioned accounts of entrapment lack the resources to account for this aspect of the moral explanation of why it is wrong to punish targets of entrapment.

The focus on the state itself, and its having compromised itself, as the main reason for why it is wrong to punish entrapped offenders also means that these views are extensionally inadequate in addition to being explanatorily incomplete. The fact that entrapment is a wrong, intimately connected with the likelihood of punishment, can make it wrong for the state to punish even when punishing 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 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 likely 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:

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

[8] 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.

> *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, one of whom is a citizen of A.

Despite the agreement, I do not think State A should punish their citizen, even though he may be fully culpable for committing the crime. Moreover, it should not punish him precisely because he was entrapped into committing it. Thus, the fact that he was entrapped provides some reason against punishing him even in cases where the punishing 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," which is conceived of as one entity that is responsible for both the entrapment and the punishment.[9] 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 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 of law enforcement. Still, some individual police officers continue to engage in it.[10] It seems morally problematic for the state to punish entrapped agents even in this scenario, but it is not clear that it is one and the same entity that is responsible for both the entrapment and the punishment. It seems to me plausible that we should place responsibility for the entrapment on the police officers and responsibility for punishment on the state itself.[11] The last worry is that several of the accounts discussed here fail to account for the fact that there are instances of morally problematic entrapment outside the legal context.[12] Consider this case:

> *Fired:* Boss A dislikes 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 to break a company rule on the job. A is watching everything on CCTV and, just as B breaks the rules, goes to fire him.

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

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

[11] One reply is that as long as "agents of the state" are responsible for the entrapment, this is sufficient to claim that "the state" is responsible for it. However, this reply is not convincing if the state has done its best to outlaw entrapment. Still, it may nevertheless be true that the state has a special responsibility for correcting the wrongs of agents of the state, which is why it should not punish the entrapped agents. 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.

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

It seems clear to me that A is trying to entrap B here. Moreover, it is because of this that it is at least *pro tanto* wrong for A to fire B (at least without a severance package) even though B has broken the rules. However, the previous accounts do not necessarily account for this. Contra the Legitimacy Account, for instance, it does not seem that A has violated a precondition for his right to fire people for fireable offences. It does not seem that bosses, in general, need to be committed to "the principles of justice" or to minimising 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 in firing B. A does not necessarily express that the offence is not fireable. First, this is because it is C who is persuading B. Second, C may not express this either – for instance, may he rely on B's dislike of the boss and the fact that the offence is an offence to motivate him to break the rules. But, as said, it seems that entrapment is still involved and, for that reason, (at least pro tanto) problematic for A to fire B.

In sum, then, although existing views capture morally salient features, I believe these worries all suggest that we also need an account of the wrongness of punishing entrapped offenders 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 these victims wrong.

III. THE CONCEPT OF ENTRAPMENT

While the motivation for the Manipulation Account comes from the 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 Gideon Yaffe's (2010) account, which is heavily inspired by Michael Bratman's (1987) theory of intention. Intentions are practical mental states that "play a role in deliberation and in the motivation and guidance of action" (Yaffe 2010, 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 one believes are incompatible with *p*. Moreover, when an intention that *p* occur plays its proper causal role, it also motivates one to pursue the actions in question. In the context of entrapment, then, we may say that the intentional condition (iii) is satisfied when the police are committed to the outcome of the target being 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 an essential feature of entrapment is that "what motivates the operation from the start is the desire to have the person convicted and punished" (2011, p. 74). Duff, Farmer, Marshall, and Tadros claim that entrapment involves inciting someone "for the purpose of arresting and prosecuting him" (2007, p. 242). Daniel J. Hill, Stephen K. McLeod and Attila Tanyi claim 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 that (iii) is essential, however. B. Grant Stitt and Gene G. James (1984) claim that there is entrapment as long as the police induce someone to commit a crime that they otherwise would not have committed.[13] But such conceptions of entrapment are mistaken. To see this, consider, for instance, the following case inspired by Tadros (2005, p. 318-9):

> *Failed Plan:* P, an undercover police officer, wants to arrest D but has no criminal evidence against him. He approaches X – who dislikes D – and encourages him to start an illegal fight with D. His plan is only to arrest D once the fight breaks out. But before he can do so another police officer arrests both X and D.

Although P satisfies conditions (i) and (ii) with respect to X – i.e., he persuaded X to commit a crime he otherwise would not have committed – and X ends up arrested, it does not seem to me that P has *entrapped* X. P does not appear to have laid out a 'trap' *for X*. He only laid one out for D. By contrast, if we assume that D's plan in persuading X to fight was to get *X* arrested, then the case is instantly recognisable as a case of entrapment. As such, we seem to have some support for the view that the intentional element is necessary for entrapment. To echo Tadros' comments on a similar case, D does not entrap X precisely because he does act "*in order to prosecute [X]*" (2005, p. 319).

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

IV. 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 seems like an *evil* or *wrongful* plan since the target is not a criminal to begin with.

[13] See also Kim (2019, p. 79-80) who also thinks there can be negligent and reckless entrapment.

Third, it is morally bad if evil or wrongful plans are successful. Fourth, therefore, it is bad if entrapped targets are punished – the evil or wrongful plans in question would be successful. The third premise is attractive but too vague.[14] 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 sense in which 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* and there is therefore a manipulation-based reason for why punishing entrapped agents is wrong. In the rest of this section, I develop these claims in more detail.

First, let me highlight that understanding entrapment involving manipulation is not entirely novel. Hughes and Kim refer to the concept in rejecting the claim that targets of entrapment can be fully responsible for their crimes. Hughes talks about manipulation-as-pressure that "undermine the autonomy of those subject to them" (2007, 58). Kim similarly says that to manipulate someone "*is* to reduce their autonomy" (2019, 75). Both authors claim, then, that the manipulation involved in entrapment diminishes the target's culpability. As explained earlier, this kind of culpability-affecting manipulation is not *essential* to entrapment. So, if entrapment always involves manipulation, it cannot be culpability-affecting manipulation.

Instead, I believe that entrapment always involves *manipulation-by-hidden-intentions*, which has previously been discussed in a more general context by Moti 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 intentions and motives. Consider:

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

Intuitively, B manipulates A. But B does not lie to, or mislead, A about the good reasons for working at D1. Perhaps he deceives her about the prospects of getting an offer from D2. If A believes that "I won't receive an offer from D2", then B might deceive A by *allowing* her to continue to have that false belief.[15] But we can stipulate this away. Suppose A and B both know there is a chance that A gets an offer from D2, but that A is ultimately persuaded by B's arguments to accept the first offer. B's behaviour still seems manipulative.

[14] Both Parr (2016) and Duus-Otterström (2017) have claimed that it is bad if immoral plans succeed. However, they do not really argue for the claim and mostly appear to accept it as a compelling intuition.

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

The best explanation for this is that B engages in *manipulation-by-hidden-intentions*. B is manipulative in persuading A because he hides his real intentions: for her to fare badly and for him to benefit. To elucidate this idea, Gorin argues that a Transparency Norm governs 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 Joseph 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 their situation. Gorin claims that, by hiding their true intentions and "playing on the expectations of manipulees […] manipulators prevent manipulees from governing themselves with an accurate understanding of their situation" (2014, p. 78). That is why hiding one's true intentions can be manipulative.

Although Gorin's claims here are attractive, it is an underdeveloped account. Claiming that we must 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:* A and B have been a couple for years and B intends to propose to A today. B wants to propose at the spot they first met, so he comes up with a fake reason for why he needs A to meet him there later today. He then proposes, and they get married soon after.

In finding an excuse to get A to go to the spot, B is hiding his real intentions. Moreover, his intention – to propose – is also relevant to A's intentions and interests. But clearly, he has not *manipulated* her.[16] Indeed, social norms seem at odds with Gorin's Transparency Norm. They push B to hide his intentions here – if not, it would ruin a great and romantic proposal.[17]

Spelling out the precise conditions for when The Transparency Norm holds is difficult, and beyond the scope of this paper. But it seems plausible that the norm at least applies when our intentions are relevant to the other person's interests *in a negative way*. That is, when hidden intentions are aimed at something bad for the other agent. If one has been pushed towards making a decision by someone who secretly wanted something bad to happen, 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.

[16] Some prefer a non-moral concept of manipulation that doesn't entail *pro tanto* wrongness. They might be happy to say that A manipulated B. But this merely pushes us to answer a different question: why is this manipulation not wrongful while B's manipulation in *Job Offer* is?

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

What is most important for our purposes is that manipulation-by-hidden-intentions is not necessarily *culpability-affecting*. Imagine, for instance, that, in *Job Offer*, it is culpable for A to decide to take the job at D1 – perhaps because it entails that she must abandon someone dependent on her staying. The culpability of that decision is not reduced by B's hiding his true intentions. It may make B extra culpable – since he is now also trying to persuade A to do something wrong – but it does not *reduce* the culpability of A'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, 379). More precisely, Raz argues for an independence condition for autonomy:

> "[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 what occurs in *Job Offer*. B pushes A towards a choice for reasons hidden from her, violating 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, B subjects A's will to his own. Her will becomes a pawn in his game – she is used as a mere means in his 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 manipulation-by-hidden-intentions. Consider the following case which is inspired by a genuine entrapment case which took place during the prohibition era in the United States:[18]

> *Sorrells*: A police officer, P, is introduced to D at D's home one night. D is a World War I veteran. P tells D that he is a World War I veteran too, and they share stories. He also tells D 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 to persuade D to buy him a drink. D finally gives in and procures him a gallon of whisky. D is then arrested by P.

P entraps D here, and this seems true regardless of the truthfulness of P's utterances. For instance, suppose that P is 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 D about his identity as a police officer, about his war experiences, or about his desire for alcohol. Still, he seems guilty of entrapping D.

This is explained by the presence of *manipulation-by-hidden-intentions*. In both versions of *Sorrells*, and in *Drugs*, it is true that the officer is hiding his true intentions in encouraging the target

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

to commit a crime. The true intention is 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 examples, what unites 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 for the existence of manipulation-by-hidden-intentions.

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 were true: if a police officer causes someone to commit a crime by acting wrongfully, it is wrong to punish the offender. It would then be wrong to punish offenders for any crime created through entrapment (i.e., manipulation). To see why 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 particular kind of wrong explains why punishing the target in the entrapment case is wrongful. Entrapment, by being an instance of manipulation which secretly intends for punishment, *aims* at punishment. It therefore morally taints the morality of punishment. The act of punishing the target is not simply the act of punishing a culpable criminal, but also the act of *completing* or *fulfilling* the wrongful manipulation which began earlier. 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.[19]

The answer is that wrongs, which are wrongs in virtue of intending for certain outcomes, are aggravated by the realisation of those outcomes. Only when they are successful or completed, by the realisation 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 where the realisation 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):

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

> *Revenge*: After their break-up, D decides to share nude photos of V, without her consent, on a "revenge porn" website.[20] A and others later view the photos on the website.

As Frowe and Parry (2019, p. 126) explain, D's wrongful act of *sharing the photos* depends on A's and/or others subsequently looking at them. The consequence of others' looking at the photos is not simply a causal consequence of D's wrongdoing but partially *constitutive* of it. Moreover, the extent to which this outcome is realised seems to make the wrong suffered by V graver – i.e., the more people look at the photos, the more seriously wrong D's initial action is. This fact also helps explain why it is wrong for A and others to look at the photos: in doing so, they enable and become complicit in D's wronging of V.

In my view, the realisation of certain consequences can likewise aggravate wrongs which are wrongs partially in virtue of *intending* that those consequences are realised. 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.[21] 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 to achieve some goal.[22] 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 a large amount of human blood to be opened. D wants to get the treasure inside. At time *t1,* she kidnaps V 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*.

D's *manipulatively harming* V is partially constituted by her intending the harm as a means to get the treasure. The gravity of that wronging can depend on consequences in the future of D's initial actions here. For instance, it can depend on the extent to which the harm-factor (i) is realised: the gravity of the wrongful drawing of V's blood seems worse the more serious side-effects V develops over time. It can also depend on the extent to which the intention-factor (ii) is realised. Suppose, for instance, that we have a chance to intervene between *t1* and *t2*. That is, we cannot prevent the kidnapping and drawing of V's blood, nor the forming of the intention to use him, but we can prevent D from getting her hands on the treasure. There is a moral reason to intervene precisely

[20] 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).

[21] See, e.g., Tadros (2011, p. 243-47; 2015).

[22] Kerstein (2013, 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.

because this will prevent V from being *successfully used* by D, which would be worse than being unsuccessfully used. In other words, the plan's success would aggravate the wronging suffered by V.

Further support for this comes from imagining a third party, A, who cannot do anything to stop D's actions but who will play some role in realising the end. For instance, suppose that A will be responsible for bringing the treasure over to D after sacrificing V's blood. Intuitively, if V knew this and could avoid being kidnapped by D by imposing some significant harm on A, he would be permitted to do so. That is, he could impose some significant defensive harm on A. The most plausible explanation for this is that A's helping realise D's intended aim would aggravate the wronging of V and therefore make A complicit in D's wronging of V.[23]

In this same way, in entrapment cases, the wrong of manipulation-by-hidden-intentions is aggravated by the extent to which the intended aim – punishment – is realised. But, of course, this is not to say that there is no wrong if the outcome is not realised. In *Enchanted Treasure*, we can blame D for a serious wrong (harming V 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 wrongdoing even if there is no punishment in the end. Still, in both these cases, the *success* of the wrong (manipulative harm or manipulation-by-hidden-intention) depends on the realisation of a certain outcome, and the realisation of those outcomes aggravates the wrongs. 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 explains the existence of a duty to refrain from punishing entrapped offenders. If realising some consequence by performing some action will aggravate the wronging of someone, then there is a duty to refrain from realising 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 constitute the realisation of the aggravating consequences, but it is obviously strongest for those who are responsible for the primary wrongful act as well. As such, we can account for the sense in which punishing entrapped agents is particularly problematic when it is one and the same state involved, but we can also account for the sense in which it is problematic to punish the entrapped agent in *Treaty*.

Moreover, we can account for what I argued was necessary earlier: 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 ultimately grounded in the same kinds of moral reasons. The second duty follows because the first is a duty that we *owe* to people. On the Manipulation Account, it is not simply the case that D is culpable, but the state has compromised the standing, authority or integrity

[23] An alternative explanation of these intuitions may be that D should not be permitted to *benefit* from her wrongdoing, which is why we should intervene, and C may be liable in virtue of helping D benefit from her wrongdoing. But this fact alone would not explain the sense that V's intervention is justified for *personal* reasons, as an act of personal resistance. Furthermore, it does not account for the fact that there is a reason to intervene also when D aim is to benefit another, innocent person.

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 account can also better account for how entrapment is problematic in general, outside of the legal context. For example, A's moral problem in *Fired* is that by actually firing B, he will fulfil the wrongful manipulation he began subjecting B to when he sent C to persuade him into committing a fireable offence.

V. OBJECTIONS

Although the Manipulation Account is a plausible theory of what is wrong about punishing entrapped offenders, it does have 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.

A. PRIVATE ENTRAPMENT

It is often said that any plausible theory of entrapment must avoid the Problem of Private Entrapment.[24] 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 a friend encouraged him to commit a crime.

On one level, the Manipulation Account does not have 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:* A loves B, but B loves D. A devises a plan to get D out of the picture by sending him to jail. He knows that D is looking at serious prison time if he is arrested now because, although he is reformed, D has been punished for several crimes in the past. A calls the police to report a crime and then begins persuading D to commit a crime. As D commits the crime, the police arrive and arrest him.

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 here, the Manipulation Account entails that D's punishment would be morally tainted. The account therefore does not completely avoid the Problem of Private Entrapment.

However, this case is not a convincing objection to the Manipulation Account. We can start by distinguishing two kinds of scenarios: one in which D is manipulated into committing a minor or victimless crime and one in which he is manipulated into committing a serious crime with genuine victims. In the latter case, there are victim-centred reason to punish D that can outweigh the entrapment-based reasons not to punish him. These may be reasons based in the in the value and

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

importance of publicly standing up for the victim and re-asserting their moral standing by holding their victimiser to account.[25] Moreover, the Manipulation Account is also compatible with holding that it is easier for the state to justify punishing D for these reasons than it would be to use the same reasons to justify punishing someone who the state has entrapped. This is because, although everyone has some reason not to fulfil or complete manipulative wrongs (by realising the intended outcomes), the duty not to do it is most stringent for the person or entity responsible for the initial wrong. This seems true of remedial duties in general. If A ends up in possession of a bike that B stole from C, then A ought to give it back. However, the duty to do so is less stringent than the duty B has to ensure that C gets the bike back. It is easier, in other words, for the state to punish D in *Envy* than it would be if the state was responsible for entrapping D.

If the crime in question did not wrong an innocent person (say, if D just delivered some illegal drugs to a consenting adult), then the Manipulation Account does seem to imply that it is wrong to punish D. But I am inclined to think that this is the right view.[26] 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 D should be punished: A would have successfully co-opted our justice system in his unjust plan to get D out of a love triangle. To put the point even stronger, there is a risk that, if we punish D, we – or the state – become *complicit* in A'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.[27]

B. MERE OPPORTUNITIES

A different objection holds that the Manipulation Account over-generalises 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 of his back pocket to lure out the pickpockets and arrest them. D grabs the wallet and tries to run away, but he is quickly apprehended.

This is proactive policing, not entrapment; most people think it is (often) permissible. 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

[25] For more expressive and communicative views on the value of punishment, see, e.g., Feinberg (1965) and Duff (2001). For more on the value and importance of communicative the moral standing of victims to wrongdoers, third parties, and victim themselves, see Statman (2008). Alm (2019) also argues in favour of reasons owed to victim to punish offenders which similarly highlights the communicative function of punishment.

[26] This is less controversial than some might think. For instance, Ho (2011, p. 92) seems open to it being wrong to punish in some instances of private entrapment. Dein and Collier (2014, p. 4) and Stark (2018, p. 8) also discuss actual cases in which a stay of prosecution was granted on the basis of private entrapment.

[27] I am grateful to an anonymous reviewer of this journal for pressing me on my original response to the issue of private entrapment.

used 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 seized the opportunity and was 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. The reason is that we should reject them.[28] 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" (1984, 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 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 defendant would have committed a crime (of similar, or more, severity) at some other stage, at which point the police would not have been able to arrest him (at least not 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. It can be permissible to manipulate and present criminal opportunities to those suspected of being criminals because they are liable to this kind of manipulation.[29]

The Preventive Test is compatible with the Manipulation Account because manipulation is not wrongful when targeted at liable agents. So, when the preventive justification is met, the manipulation involved in the proactive policing will not be wrongful. It is therefore also permissible to punish the target of the manipulation. For that reason, the Manipulation Account is probably 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 implication is intuitively correct.

C. 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 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 jail). But consider:

[28] This is not a novel position. Stitt and James (1984), Dworkin (1985), Howard (2016) and Lippke (2017) are all sceptical of both tests.

[29] I borrow from Nathan's (2017) argument that people engaged in criminal activity are liable to sting operations and therefore not wronged by the deception and manipulation often involved in sting operations.

> *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 for D to be punished *if* he 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 but hopes 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 entrapment is problematic precisely because it is a form of virtue testing.[30] They may also insist that it is wrong to punish in *Virtue Testing*.

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

The second response is to accommodate the claim that it is wrong to punish in *Virtue Testing*. In contrast to *Failed Plan*, *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:

> *Carjacking*: D approaches V in her car and threatens to shoot her unless she hands over her keys. V complies and runs away. Later, D is arrested and testifies that he did not intend to use the gun if V complied but that he would have done so if she had resisted.[31]

Is Blake guilty of the U.S. federal crime of carjacking *with the intent to cause death or serious bodily harm*? In one sense, no. He only *conditionally* intended to cause harm. His main plan was simply to scare Violet into handing over the keys. But suppose we add that Blake was confident that Violet would resist and so anticipated that the condition in his conditional intention would be satisfied. In that case, we might lean more towards finding him guilty.

Yaffe (2004) provides a helpful way of thinking about these cases. Recall, on his view, intentions are tied to various rational commitments that structure one's deliberations. If you intend to *x* soon, then you are rationally committed to various things: to pursue actions and steps essential to *x*ing, not to form intentions to do things incompatible with those actions, and so on. To simplify a bit, Yaffe builds on this and suggests that we look at how the conditional intention structured Blake's deliberations by comparing the structure of his deliberations to the deliberations of those

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

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

who carjack with the intention to harm and of those who carjack without such an intention. For instance, if Blake had spent little time contemplating whether Violet would resist, had not been careful to check if the gun was actually loaded, and so on, then we may conclude that he is guided by deliberations more similar to someone who carjacks without an intention to cause harm. However, if he thought resistance was likely, made sure the gun was loaded, had already bought supplies to clean blood spills from the interior, and so on, then he seems guided by deliberations more similar to someone who carjacks with the intent to harm. In the first case, we might reasonably find him not guilty of the federal crime, but in the second, we might reasonably find him guilty

This provides one way of understanding "purely" virtue testing entrapment cases. Realistically, many such officers, like P, are likely to be guided by deliberations similar to those of paradigmatic entrapping officers. For instance, by trying to get D to commit the crime, he knows that he is increasing the likelihood of it happening. If he is serious about his conditional intention to arrest him, he is likely also to have deliberated about how to arrest him, how to prevent escape, made sure there is room for a suspect in the police car, and so on. In that case, treating *Virtue Testing* as equivalent to entrapment unconditionally aimed at punishment may be appropriate.

Compare this to a version of *Virtue Testing* in which P is guided more by his hope that D does not commit the crime. For instance, he has not made concrete plans for arresting him if the crime is committed. Imagine, indeed, that P does not want D to be punished, but that he knows he has to do his job of trying to arrest those who commit crimes. So, to increase the likelihood that D will go unpunished, he leaves his gun at home, knowingly making it easier for D to escape if P tries to arrest him. To his surprise, when D gives in to the virtue testing, another police offer steps in and arrests him. Here, although there is a conditional intention to have him arrested and punished, I am much less inclined to think it is an instance of entrapment which morally taints punishment.

Ultimately, *Virtue Testing* is not a counterexample to The Manipulation Account. There are ways for the account to treat some of these cases as equivalent to paradigmatic entrapment cases. Furthermore, those which cannot be treated as similar in this way do not seem to be relevantly similar regarding whether punishment is morally tainted either.[32]

VI. CONCLUSION

Entrapment is puzzling. The puzzle is generated by seemingly conflicting intuitions concerning the wrongness of entrapment and the wrongness of punishment on one side and the culpability of the offender on the other side. 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 can account for the sense in which it is inherently problematic to punish those who are victims of entrapment. Punishing an entrapped offender is to fulfil or complete the wrongful manipulation

[32] 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.

they are victims of. This is precisely why there is also a pro tanto reason *not* to impose punishment on them, despite them being culpable criminals.

BIBLIOGRAPHY

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: 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*, Volume 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. Coons, C. 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.

Pettit, P. 2018. Three Mistakes About Doing Good (and Bad). *Journal of Applied Philosophy* 35: 1-25.

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*, eds. Oberdiek, J. 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. White, M., 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.