Law and Philosophy (2017) 36: 707–750
DOI 10.1007/s10982-017-9301-9

ALEXANDER SARCH

WHO CARES WHAT YOU THINK? CRIMINAL CULPABILITY AND THE IRRELEVANCE OF UNMANIFESTED MENTAL STATES

(Accepted 10 April 2017)

ABSTRACT. The criminal law declines to punish merely for bad attitudes that are not properly manifested in action. One might try to explain this on practical grounds, but these attempts do not justify the law's commitment to *never* punishing unmanifested mental states in worlds relevantly similar to ours. Instead, a principled explanation is needed. A more promising explanation thus is that one cannot be *criminally culpable* merely for unmanifested bad attitudes. However, the leading theory of criminal culpability has trouble making good on this claim. This is the theory that an action is criminally culpable to the extent that it manifests insufficient regard for legally protected interests. The trouble is that this theory's defenders have not adequately explained what it is for an action to *manifest* insufficient regard. In this paper, I aim to provide the required account of manifestation, thereby rendering the insufficient regard theory more defensible. This, in turn, allows the view to explain the broad range of doctrines that treat unmanifested mental states as irrelevant. The resulting theory of criminal culpability is both descriptively plausible and normatively attractive. Moreover, it highlights the continuity between criminal culpability and moral blameworthiness by showing how the former functions as a stripped-down analogue of the latter.

Many criminal law doctrines treat mental states that are not manifested in conduct in the right way as irrelevant. A 'basic premise of Anglo-American criminal law is that no crime can be committed by bad thoughts alone'.[1] Nor do we enhance punishments merely because we know one was willing to offend in worse ways. If I want to kill my enemy

[1] Wayne LaFave, 1 Subst. Crim. L. § 6.1 (2d ed.). *See also* Michelson v. United States, 335 U.S. 469, 489 (1948) (Rutledge, J., dissenting) ('Our whole tradition is that a man can be punished … only for specific acts defined beforehand to be criminal, not for general misconduct or bearing a reputation for such misconduct'.); Kenneth W. Simons, *Does Punishment for 'Culpable Indifference' Simply Punish for 'Bad Character'? Examining the Requisite Connection Between Mens Rea and Actus Reus*, 6 Buff. Crim. L. Rev. 219, 234 (2002) ('the harsh sanctions of the criminal law should not be brought to bear on individuals who have not yet done anything wrong, but who merely have disreputable – or even dangerous – character traits').

but am bribed to only beat him up instead, I'm not guilty of murder simply because I was willing to go that far. As Ken Simons notes, 'we are … properly[] reluctant to impose punishment on a person simply for [attitudes or characteristics] unless and until [they] are expressed in action'.[2] Similar thinking lies behind the concurrence requirement between mens rea and actus reus: the former must be expressed in (i.e. cause) the latter in the right way for there to be a proper basis for guilt.[3]

Furthermore, one's *motives* for breaking the law generally do not matter.[4] A thief is not guilty of a more serious offense because he is motivated by hatred of the victim rather than wanting to be able to afford a better school for his child. In determining criminal guilt, such motives don't count as manifested. The affirmative defenses function similarly: '[W]hen an individual finds himself in a position where the law grants him the right to kill another in his own defense, it makes no difference whether his dominant motive is other than self-preservation'.[5] An executioner authorized to carry out a death sentence commits no crime even if his actual motivation is the desire to see the prisoner die gruesomely.[6] However, there are exceptions to the criminal law's general lack of concern with motives. A few offenses – like treason,[7] kidnapping[8] and hate crimes[9] – include bad motives as an element. Motives may also affect sentencing.[10]

---

[2] Simons, *supra* note 1 at 234.

[3] Under this rule, guilt requires not only that one possesses the required mens rea while doing the relevant actus reus, but also that 'the defendant's mental state actuate[] the physical conduct'. WAYNE LAFAVE, 1 SUBSTANTIVE CRIMINAL LAW § 6.3 (2d ed.). *See also* JOSHUA DRESSLER, UNDERSTANDING CRIMINAL LAW 199 (6th ed.) ('[t]he defendant's conduct that caused the social harm must have been *set into motion or impelled by* the thought process that constituted the *mens rea* of the offense'); Alex Sarch, *Knowledge, Recklessness and the Connection Requirement Between Actus Reus and Mens Rea*, 120 PENN. ST. L. REV. 1 (2015).

[4] WAYNE LAFAVE, 1 SUBST. CRIM. L. § 5.3 (2d ed.) ('[M]otive, if narrowly defined to exclude recognized defenses and the 'specific intent' requirements of some crimes, is not relevant on the substantive side of the criminal law'.).

[5] *Id.*

[6] *Id.*

[7] Treason requires purpose to aid an enemy of the state, and mere knowledge that this will result does not suffice. Haupt v. United States, 330 U.S. 631, 641–42 (1947); WAYNE LAFAVE, 1 SUBST. CRIM. L. § 5.2 n.9.

[8] Kidnapping often requires a prohibited purpose like obtaining a ransom or terrorizing the victim. *See* MODEL PENAL CODE § 212.1 (stating that one 'is guilty of kidnapping if he unlawfully removes another … with any of the following purposes: (a) to hold for ransom or reward, or as a shield or hostage; or (b) to facilitate commission of any felony or flight thereafter; or (c) to inflict bodily injury on or to terrorize the victim or another').

[9] Hate crimes carry harsher punishments if one acted 'because of the actual or perceived race, color, religion, or national origin of any person'. 18 U.S.C. § 249.

[10] *See* Carissa Byrne Hessick, *Motive's Role in Criminal Punishment*, 80 S. CAL. L. REV. 89, 90 (2006).

My aim here is to explain this jumble of doctrines concerning unmanifested mental states. I won't do so by offering a general theory of what it is for an action to manifest a mental state. Rather, I defend a theory of criminal culpability that explains why culpability is not impacted by the mental states that criminal law doctrine treats as unmanifested.

My topic therefore is *criminal culpability*, not moral blameworthiness.[11] The latter, roughly, is what makes one an apt target of reactive emotions like resentment and indignation, while the former determines the extent to which one's conduct merits condemnation by the criminal law.[12] All else equal, criminally culpable act tokens warrant punishment and act types that comprise solely such tokens are aptly criminalized (perhaps on the condition that it's not bad policy).

My theory follows the growing trend of understanding criminal culpability in terms of insufficient regard. On this view, one is culpable for an action to the extent it manifests insufficient regard (or disrespect) for the legally protected interests of others or protected values more generally. Insufficient regard is a species of ill will – more specifically, making mistakes in the way one recognizes, weighs and responds to the reasons that bear on how to act.[13] If one attaches so little weight to others' interests that one burns down a building for the insurance despite knowing someone is inside, one acts from insufficient regard. Larry Alexander and Kim Ferzan endorse such a view,[14] as do Peter Westen,[15] Gideon Yaffe[16] and others.[17]

One attraction of this view is that it both preserves a connection to moral blameworthiness and recognizes the ways in which the law is different. On the influential quality of will theory, acts are morally

[11] In principle these concepts could be identical, though I mean to leave it open precisely how they're related.

[12] At least this is the case setting aside considerations of luck. Some might think the punishment one merits can be impacted not only by culpability, but also by how much harm one causes – even if due in part to luck. (Thanks to Massimo Renzo for this worry.) I doubt this would be normatively defensible, but can't argue for this view here. Rather, I aim to sidestep the issue by focusing on pairs of defendants who both cause the same amount of harm.

[13] Gideon Yaffe, Attempts 38 (2011) (endorsing the view that an action is culpable to the degree that 'it is a product of a faulty mode of recognition or response to reasons for action').

[14] Larry Alexander and Kimberly Ferzan, Crime and Culpability 67–68 (2009) (arguing that 'insufficient concern [is] the essence of culpability').

[15] Peter Westen, *An Attitudinal Theory of Excuse*, 25 Law & Philosophy 289, 373–74 (2006) ('a person is normatively blameworthy for engaging in conduct that a statute prohibits if he was motivated by an attitude of disrespect for the interests that the statute seeks to protect').

[16] *See* Yaffe, *supra* note 13; Gideon Yaffe, *Intoxication, Recklessness, and Negligence*, 9 Ohio St. J. Crim. L. 545, 552–53 (2012). *See also* Yaffe *infra* notes 21 and 54.

[17] Simons, *supra* 1 note at 249–50; Alex Sarch, *Double Effect and the Criminal Law,* Crim. L. & Philosophy (2015).

blameworthy to the extent they manifest ill will, which can be understood as attaching insufficient weight to the relevant *moral reasons*.[18] The insufficient regard theory takes criminal culpability to consist in manifesting insufficient regard for the applicable *legally recognized* reasons. Thus, criminal culpability would be a simplified analog of moral blameworthiness, with the former sensitive to only some considerations that impact the latter.

Despite its attractions, the insufficient regard theory of criminal culpability also faces challenges. Most importantly, it is far from clear what it means for an action to *manifest* insufficient regard. Proponents of the view are rarely precise about how to determine the *amount* of insufficient regard an act manifests.[19] But the theory cannot do without an account of this key notion, given the many doctrines showing that punishment does not directly correspond to how much insufficient regard one *possesses* but does not manifest. One might possess tremendous ill will towards others, but if it's not revealed in action in the right way, the appropriate predicate for punishment is not present. What I'll argue is that answering this challenge for the insufficient regard theory lets us account for the above doctrines concerning unmanifested mental states.

Of course, it might be tempting to explain these doctrines merely as a function of our epistemic and practical limitations. It will often be complicated and costly to determine and evaluate the attitudes the defendant had while acting. So perhaps it is prudent for the law to ignore unmanifested bad attitudes.[20] However, this pragmatic

[18] *See* Julia Markowitz, *Acting for the Right Reasons*, 119 PHIL. REV. 201 (2010); NOMY ARPALY AND TIM SCHROEDER, IN PRAISE OF DESIRE 170 (2014).

[19] Throughout this paper, I assume that insufficient regard is a scalar concept. This is because I take criminal culpability itself to be scalar. As I understand it, culpability is the main – perhaps only – determinant of how serious an offense one is guilty of and thus the sentencing range one is subject to. Thus, if the present theory is to capture this scalar notion of criminal culpability, insufficient regard must itself be taken to come in degrees.

[20] This position has recent defenders. *See, e.g.*, David Lefkowitz, *Blame and the Criminal Law*, 6 JURISPRUDENCE 451, 462–65 (2015). Lefkowitz argues it's for practical reasons that we shouldn't punish unmanifested insufficient regard, even though in principle we could. He mentions three practical limitations. First, there is an 'epistemic challenge' in proving that one who complies with the law 'nevertheless lacks an attitude of proper regard for others' legally protected interests'. *Id*. at 465. Second, we face the 'practical challenge of designing and operating a criminal justice system that punishes … mere bad attitude crimes without [causing] corruption and abuse'. Third, the state's limited resources might prevent it from 'even investigating your bad attitude crimes, let alone prosecuting them'. *Id*. *See also* Martin R. Gardner, *The Mens Rea Enigma: Observations on the Role of Motive in the Criminal Law Past and Present*, 1993 UTAH L. REV. 635, 685–86 (1993) (using practical reasons to show why the law shouldn't require proof of 'evil motive': '[a]ctual motives are often hidden … in the subconscious' and 'are difficult to evaluate'; and '[s]erious … motivational analysis would require trial courts to consider detailed case histories of each defendant', which often are 'unavailable to prosecutors at the time necessary to charge the crime').

explanation does not capture the full strength of the principles the law is committed to. Suppose our practical limitations are temporarily alleviated. Most simply, the defendant might simply *admit* (credibly) that he had very bad attitudes while acting. Even then, however, it is doubtful that we should punish him merely for his bad attitudes or do so more harshly for his willingness to commit worse crimes. Still, in such a case, our practical limitations would provide no bar to doing so.

Thus, my aim is to offer a *principled* explanation of when and why criminal law doctrine is not concerned with unmanifested mental states. I'll offer an account of what it means to manifest a particular *level of insufficient regard*, which explains (i) why one is not culpable for bad attitudes one merely possesses but does not act on, and (ii) why one is not more culpable for breaking (or complying with) the law for bad rather than good motives. In this way, answering the main challenge for the insufficient regard theory endows it with significant explanatory power. The resulting theory, I'll argue, is highly normatively attractive as well.

The paper proceeds as follows. Section I recounts the reasons why the insufficient regard theory must include a manifestation requirement at all. Section II argues against the two most natural accounts of manifesting a given level of insufficient regard: the *purely causal account*, and the *epistemic account*. Section III then presents my own account. It is partly causal, but also draws on what Gideon Yaffe has dubbed a 'principle of lenity'.[21] While Section III aims to give a *descriptively adequate* framework, Section IV sketches a normative argument for the view.

I. WHY A MANIFESTATION REQUIREMENT IS NEEDED

A theory of criminal culpability should be *descriptively adequate* in the sense that it captures the core features of the law in the jurisdictions one cares about. My focus is Anglo-American criminal law. To capture its core features, the insufficient regard theory needs a manifestation requirement. This can be seen by considering some unsatisfactory formulations of the theory.

[21] Gideon Yaffe, *The Point of Mens Rea: The Case of Willful Ignorance*, CRIM. L. & PHILOSOPHY (2016).

To start, our theory must respect the voluntary act requirement. This is the rule that we don't punish without voluntary conduct.[22] If one unconsciously lashes out while asleep, one is not guilty of assaulting one's partner. One has no criminal culpability. This suggests:

(IR*): The criminal culpability of D's voluntary action A equals the amount of insufficient regard D possesses while doing A.

However, the voluntary act requirement is notoriously easy to satisfy. As long as the defendant engaged in *some* relevant voluntary conduct, the requirement is satisfied.[23] Thus, (IR*) is inadequate. It allows one to be culpable merely for innocuous actions done while possessing a deficient level of regard for others. Consider:

TED THE WOULD-BE TERRORIST (V.1): Ted hates people he perceives as ''foreigners'' and wants to bomb a mosque. But he hasn't taken any steps towards doing so yet. A CIA agent learns of his sentiments and bribes him $10,000 not to bomb the mosque. Ted agrees. He is willing to go through with the crime and the only reason he refrains is the bribe. Instead, Ted decides to paint his fence – though while he does so his hatred of Muslims has not abated.

Ted's act of painting the fence satisfies the voluntary act requirement, and (IR*) entails that he is criminally culpable. But this is implausible. While Ted is clearly highly morally blameworthy and has a deplorable character, he is not appropriately subject to criminal liability. We do not punish merely for a willingness to offend. As LaFave puts it, a 'basic premise of Anglo-American criminal law is that no crime can be committed by bad thoughts alone'.[24]

Perhaps Ted is not criminally culpable because he did not violate any legal prohibition of the sort that we'd likely see in Anglo-American jurisdictions. Thus, a better theory would be:

[22] *See*, *e.g.*, MODEL PENAL CODE § 2.01(1) ('A person is not guilty of an offense unless his liability is based on conduct that includes a voluntary act or the omission to perform an act of which he is physically capable'.); LaFave, *supra* note 1 (observing that criminal liability requires 'an act, or an omission to act where there is a legal duty to act', and that a 'bodily movement, to qualify as an act forming the basis of criminal liability, must be voluntary').

[23] To see how easily satisfied this requirement is, note that one can be guilty of criminal homicide despite being asleep at the time one causes death if one previously felt extremely sleepy but continued to drive anyway. Here, the defendant's 'voluntary act consists of driving the car, and if the necessary mental state can be established at the time' (e.g. recklessness) it is enough for guilt. LaFave, *supra* note 1. Moreover, even possession can count as an act. MODEL PENAL CODE § 2.01(4).

[24] LAFAVE, *supra* note 1.

(IR**): D's action A is criminally culpable to degree *n* if and only if (1) A violates a criminal prohibition and (2) A was done while D possessed a level of insufficient regard that equals *n*.

However, this will not do either. For one, (IR**) entails that actions can't be criminally culpable without violating existing criminal prohibitions. But some act types may not yet have been criminalized. So (IR**) can at best only capture jurisdiction-specific culpability attributions, not the underlying notion of culpability that the law *should* be tracking. Our theory should do both.

Moreover, (IR**) is implausible because it attributes enhanced culpability to defendants who are willing to behave in worse ways than they did. Suppose Ted *did* violate a legal prohibition:

TED THE WOULD-BE TERRORIST (v.2): Ted still is committed to bombing a mosque. But he can't get ahold of any explosives right now. Instead, he does the greatest amount of damage he is able to under the circumstances, which is to spray-paint anti-Muslim slurs on a public building – a criminal offense where he lives.

(IR**) entails that Ted is highly culpable. He engaged in voluntary conduct that violates a criminal prohibition while possessing seriously deficient regard for others. According to (IR**), Ted's culpability is pegged to what he is willing to do (i.e. attempt a bombing), and thus is far greater than that of someone who paints racist slurs *without* being willing to bomb a mosque. But this is implausible. Just as we do not punish for a bare willingness to offend, we also refrain from punishing merely for a willingness to behave in worse ways than one actually did. As Simons notes, 'we are … properly[] reluctant to impose punishment on a person simply for [attitudes or characteristics] unless and until [they] are expressed in action'.[25] The full level of insufficient regard Ted possesses is not completely manifested. He seems less culpable than (IR**) entails.

To avoid these problems, we need a manifestation requirement. Thus, the theory should say:

(IR): D's action A is criminally culpable to degree *n* if and only if the degree to which A *manifests* insufficient regard equals *n* (regardless of how much D happened to possess when doing A).

[25] Simons, *supra* 1 note at 234.

(IR) matches legal theorists' canonical statements of the theory,[26] and helps explain why we don't enhance punishments merely because one was willing to offend in worse ways than one actually did. After all, without actually doing the worse offense, one's willingness is not manifested. But what, in turn, explains *this*? (IR) remains incomplete because we still need to know what it *is* to manifest insufficient regard to a particular degree.

II. EXPLAINING MANIFESTATION IS HARDER THAN YOU MIGHT THINK

Explaining how much insufficient regard an action manifests is the big challenge for such theories. This section argues against the two most natural proposals.[27] I do not claim to conclusively refute these views, but merely seek to motivate my own account presented below.

A. The Pure Causal Approach

The most natural account takes it that an action manifests a given amount of insufficient regard just in case possessing this amount is what *causes* one to do that act. This approach is naturally suggested by the *concurrence requirement*, which requires that the mens rea of the offense must 'actuate' the requisite actus reus.[28] This requirement embodies a theory of mental state expression: An action expresses a mental state, intention or attitude just in case the former causes the latter (in the right way). However, what we want is an account of manifesting a given level of *insufficient regard*, which depends on one's whole configuration of attitudes and reflects how one weighs the relevant reasons. Thus, the causal theory we're interested in is this:

Pure causal approach: An action A manifests a level of insufficient regard *n* if and only if (and because) A is caused by a configuration of mental states that constitutes level *n* of insufficient regard towards legally protected interests or other protected values.

---

[26] *See supra* notes 14–16.

[27] Elsewhere I have argued against other attempted solutions to this problem, particularly as it arises for moral blameworthiness. *See* Alex Sarch, *Review: N. Arpaly & T. Schroeder, In Praise of Desire (Oxford 2014)*, 31 Economics & Philosophy 320, 324–27 (2015).

[28] *See supra*, note 3.

I call this the *pure* causal approach, since it might just be one component of a more sophisticated theory of manifestation. In fact, this is true of my account, which is a *partly* causal view.

The pure causal approach is attractive, since it captures the insight behind the concurrence requirement. Unless this requirement is satisfied (e.g. if one only acquires the mens rea of the crime *after* doing the actus reus), the pure causal approach does not permit the full degree of insufficient regard inherent in the crime to be manifested in one's conduct.

However, the pure causal approach has problems. Simply put, causation is not *sufficient* for manifestation. As many cases show, the amount of insufficient regard that causes an action does not always correspond to the amount the action manifests. Recall the following chestnut[29]:

KILL YOUR UNCLE: Dennis and Charlie intend to kill their respective uncles today. They both have exactly the same motivations and attitudes. Both take a drive to clear their heads and think further about when to do the killing. While driving, Dennis happens to see his uncle on the street and shoots him then and there. Since Dennis intentionally kills, he is a murderer. By contrast, Charlie, while lost in thought and driving carelessly, hits a pedestrian, who turns out to be his uncle – his intended victim. Charlie, unlike Dennis, did not commit murder. He caused his uncle's death, but only negligently (not intentionally).

Dennis and Charlie are stipulated to have exactly the same motivations and attitudes.[30] They both intend to kill. They have equally deplorable characters and levels regard. On some views, they would be equally morally blameworthy.[31] Still, there is an intuitive sense in which what Charlie *did* is less culpable. Dennis is a murderer, but we wouldn't affix that label to Charlie.

The insufficient regard theory would try to capture this by maintaining that Charlie's conduct did not *manifest* as much insufficient regard as Dennis's did, such that Charlie's conduct is less culpable. But it's unclear whether the pure causal approach to manifestation can deliver this result. Charlie's high level of insuffi-

[29] I've adapted this example from Simons's discussion of the case, which he culls from John Searle. *See* Simons, *supra* note 17 at 232; JOHN SEARLE, INTENTIONALITY: AN ESSAY IN THE PHILOSOPHY OF MIND 82 (1983).

[30] As a reviewer notes, we must suppose neither Charlie's nor Dennis's plan includes a conditional intention to kill their uncle through careless driving if the opportunity presents itself. This might also make Charlie guilty of murder.

[31] P. Graham, *A Sketch of a Theory of Blameworthiness*, 88 PHIL. & PHENOMENOLOGICAL AFF. 388, 396–99 (2014).

cient regard – i.e. the configuration of attitudes that includes his intention to kill – does seem *somewhat* causally active in producing the conduct that resulted in his uncle's death. It caused him to take the drive. This suggests that the amount of insufficient regard manifested is not identical to that which is causally active in producing one's conduct. (There are answers a proponent of the pure causal approach may try, but I doubt they succeed.[32])

As noted below, this case might be handled by other theories (like Alexander and Ferzan's). Still, the present problem goes deeper. Other cases show that an action might not manifest all the insufficient regard that caused it. Consider a legal version of a case of Arpaly and Schroeder's[33]:

TRIVIAL WRONGS: Jack and Jill each obtain software that removes fractions of cents from others' bank accounts and transfers the money to the software operator. Jack decides to use the software to enrich himself because he naively sees it as a practically harmless prank. He would not be willing to impose greater harms even if he could. By contrast, Jill harbors profound hatred towards the members of her community (whom she takes to be sinful and unclean). She wants to impose as much pain and suffering on them as possible. It just so happens that this is the greatest harm she is currently able to inflict since she is in a remote location and very sick. Given her bottomless ill will, she would jump at the chance to cause vastly more harm (even death) if she could. Indeed, she hopes to use the money obtained with this software to some day purchase the means to wreaking more havoc (though she hasn't decided how). Jack and Jill's conduct is otherwise the same: They each use the software to remove a total of $500 from a large number of bank accounts.

Jill's conduct is caused by vastly more insufficient regard than that which caused Jack to act the same way. Thus, the pure causal account entails that Jill's conduct is vastly more culpable than Jack's. But this is doubtful, I submit. Both impose the same small harms on their victims and neither actor has any justification. Instead, their

[32] A proponent of the pure causal approach might respond that Charlie's killing his uncle was caused only by his mental state of *negligence*, not the further mental state of intent to kill. Perhaps Charlie's intention to kill was merely free-floating – i.e. not causally active in producing the conduct that led to the death. However, this response requires the questionable assumption that Charlie's configuration of mental states at the time can be carved up into different components, one of which – his negligence – was causally active, and another of which – his intent to kill – was not. However, it is not clear that mental states can be separated so neatly. Wasn't Charlie's intent *somewhat* causally active? He went for a drive to plan the killing, so he was driving in part because of his intention. Perhaps, then, the intention didn't play *enough* of a causal role in causing the act. Still, it is extremely hard to specify how much of a causal role a mental state must play for it to impact the insufficient regard an action manifests.

[33] *See* Arpaly and Schroeder, *supra* note 18 at 188–89 (discussing a praiseworthiness analog about a lost motorist).

acts seem roughly equally culpable: They would be convicted of equally serious offenses and face the same sentencing range.[34] Accordingly, the full degree of insufficient regard that caused Jill's conduct does not seem to be *manifested* in her offense. What cases like this show, as Arpaly and Schroeder note, is that the amount of 'ill will that an action manifests is not the same as the amount of … ill will that exists and is being acted on'[35] – a point the pure causal approach fails to capture. (Arpaly and Schroeder suggest that only some acts provide *occasion* to manifest extreme ill will, but I've argued this reply fails.[36])

A final, related problem is that the pure causal approach has trouble capturing the default rule that motives (i.e. one's aims in acting) generally are 'not relevant on the substantive side of the criminal law'.[37] This rule is familiar, e.g., from the law's approach to defenses. Consider:

EVIL TROLLEY TURNER: Darryl finds himself in the classic trolley scenario. A trolley is careening toward five people who are tied to the tracks. It will kill them unless Darryl pulls the switch, and diverts the trolley onto another track to which just one person is tied, Vicky. Thus, if Darryl pulls the switch, only one person will die, rather than five. Darryl knows all this and decides to divert the trolley onto the other track. However, his sole reason for doing so is that he hates Vicky and wants her dead. He doesn't care one whit for the five on the other track, and he wouldn't bother to pull the level unless doing so would allow him to accomplish his aim of killing Vicky.

Darryl is highly morally blameworthy. He has a horrible character, and harbors deplorable attitudes and is willing to act on them. Nonetheless, in many jurisdictions, Darryl would get the benefit of a necessity defense. As LaFave notes, when 'the law grants [a person] the right to kill another in his own [or someone else's] defense, it makes no difference whether his dominant motive is other than self-preservation'.[38] '[T]he law is not concerned with motive once facts

[34] An anonymous reviewer notes that even if Jack and Jill face the same sentencing range, perhaps Jill should be sentenced at the high end of the range and Jack at the low end. Still, the pure causal approach is in trouble for taking Jill to be *vastly* more culpable than Jack. In addition, even if Jack and Jill should be sentenced slightly differently, I suspect that this is due to the broader range of considerations that legitimately impact sentencing – e.g. deterrence and rehabilitation. It does not obviously show a difference in culpability of the sort that matters at the guilt stage of a case. Moreover, I am open to the idea that sentencing can consider more fine-grained differences in moral blameworthiness than typically would (or should) matter at the guilt stage. (I discuss this at length in Subsection III.C.)

[35] *Id.*

[36] Sarch, *supra* note at 27 at 324–27.

[37] WAYNE LAFAVE, 1 SUBSTANTIVE CRIMINAL LAW § 5.3 (2d ed.).

[38] *Id.*

supporting the defense have been established'.[39] Many legal scholars endorse this view.[40]

Nonetheless, the purely causal approach cannot deliver this result. Darryl was caused to act as he did by a configuration of attitudes that constitute an extremely high level of disregard for the protected interests of others. Thus, this approach entails Darryl is highly criminally culpable.

Of course, some might challenge the criminal law's position on this issue by trying to individuate actions more finely. One might claim that (i) switching to kill Vicky is a different action from (ii) switching the trolley to save the five. If acts can be individuated this finely, then we might say that what Darryl did was the highly culpable act of doing *(i) rather than (ii)*. Of course, this strategy faces difficult questions about how to describe and individuate actions.

While I take this response to be a live option (it has well-known defenders[41]), I want to offer a theory of criminal culpability that at least is *able* to capture the criminal law's approach from above, which is maximally generous to criminal defendants. That is, I want to be able to explain why it at least *makes sense* for defendants to get the benefit of all the defenses available to them based on the facts they were aware of in acting. The view I defend below can do this by taking it that Darryl's bad attitudes do not count as *manifested*.

Notice that the default rule that motives don't matter is not just a feature of the way the law treats defenses. An analogous point applies with respect to the elements of a crime. Consider:

GOOD THIEF, BAD THIEF: Mary's child is having a hard time at pubic school, so she steals $5000 from an ATM to be able to send her child to a better private school. By contrast, Barry is bored and wants to go to Vegas. He steals $5000 from an ATM to be able to afford the trip.

The pure causal approach says that because Barry's action was caused by a configuration of mental states that constitutes a much more deficient level of regard than that which caused Mary to act, Barry is more criminally culpable. But this fits poorly with legal

[39] *Id.*

[40] Larry Alexander, *The Means Principle*, in LEGAL, MORAL, AND METAPHYSICAL TRUTHS: THE PHILOSOPHY OF MICHAEL MOORE 28 (K.K. Ferzan and S.J. Morse, eds., 2014) ('[I]f the switcher believes switching will save the five, then his switching the trolley is nonculpable *even if his only reason for doing so is to kill the one on the siding*'. (emphasis in original)). *See also* Alexander & Ferzan, *supra* note 14 at 60–61 (defending a similar view).

[41] *See* VICTOR TADROS, THE ENDS OF HARM 156 (2011) (endorsing this response to the present case).

practice. In most jurisdictions, Barry would not be treated as more criminally culpable and convicted of a worse offense just because he had worse motives for breaking the law.[42] Neither Mary nor Barry has a recognized justification for their conduct.[43] Granted, Barry's worse motives might matter at sentencing, but at least when it comes to the offense he is to be convicted of – i.e. the level of culpability imputed to him at the guilt stage of the trial, which fixes the applicable range of sentences – Barry's worse motives are irrelevant. Thus, we want our theory to explain why motives do not matter to the sort of criminal culpability that is central to substantive criminal law doctrine applicable at the guilt stage. The pure causal theory isn't up to the job.[44]

B. Other Theories?

My present aim is to motivate the account of culpability I offer below. However, as a reviewer pointed out, I won't succeed in this if other theories easily capture the cases I discuss. I cannot argue against all other theories here. But to further motivate my view, let me note some challenges for one other leading insufficient regard theory, which also is causal in nature.

Alexander and Ferzan take culpability to be a function both of (i) the magnitude of risk that one perceives one unleashes on others and (ii) the reasons that impelled one to do so.[45] Keeping the one factor

---

[42] *See* WAYNE LAFAVE, 1 SUBST. CRIM. L. § 1.2 (2d ed.) ('[A] good motive will not normally prevent what is otherwise criminal from being a crime. Thus it is nonetheless … larceny that one stole a rich man's money to give his impoverished family a better life'). *See also U.S. v. Badolato*, 701 F.2d 915, 921 (11th Cir. 1983) (stating that for crimes aimed at financial gain, 'a person would be just as guilty if the motive were … to obtain money for a worthy cause').

[43] WAYNE LAFAVE, 1 SUBST. CRIM. L. § 5.3 (2d ed.) ('[T]here are numerous … cases where a person has been found guilty of a crime in spite of what might be viewed as a good motive in committing it').

[44] A reviewer suggests the pure causal approach might be modified to say that an act is criminally culpable to degree *n* if it is *proximately* caused by a degree of insufficient regard equal to *n*. If proximate cause is understood in terms of directness, I don't think this will help. In TRIVIAL WRONGS, Jill's minor misconduct is directly caused by a *tremendous* degree of insufficient regard. Nonetheless, her act doesn't seem tremendously culpable. It is only roughly as culpable as Jack's. Thus, it might be better to understand proximate cause in terms of normative considerations designed to limit the implausible implications of the view. However, in that case, the theory begins to look indistinguishable from the normatively constrained causal theory I defend below in Section III. On my theory, the amount of insufficient regard manifested is determined using normative assumptions like the principle of lenity. Thus, my theory is partly normative just like the best understanding of the proximate cause approach. Still, my theory is more illuminating because it specifies exactly *how* the relevant normative considerations operate, rather than letting them be diffusely applied through intuitions about what suffices for proximate causation.

[45] *See* Alexander & Ferzan *supra* note 14 at 18.

fixed, one is more culpable the worse one is with respect to the other factor – i.e. the greater the perceived risks one unleashes or the worse the reasons that led one to do so. The theory is causal in nature because it makes culpability depend in part on the reasons that caused one to act, which reflect one's level of regard for others.

Alexander and Ferzan can capture KILL YOUR UNCLE. An actor, they say, is not culpable 'until he engages in conduct that he believes unleashes a risk of harm over which he no longer has complete control'.[46] Charlie does not irrevocably unleash as great a risk of death as Dennis, so Charlie is less culpable than Dennis although they acted for equally bad reasons.

Nonetheless, Alexander and Ferzan's view does not sit comfortably with the default rule, seen above, that motives generally don't matter in substantive criminal law doctrine. It is at least a point in favor of a theory if it gives a principled explanation of how this default rule can be sensible. In TRIVIAL WRONGS, Jack and Jill irrevocably unleash the same magnitude of perceived risk on others. Because Jill does so for vastly worse reasons than Jack, on natural assumptions it follows from Alexander and Ferzan's view that Jill's act is vastly more culpable than Jack's.[47] Nonetheless, it fits better with legal practice to take the two to be roughly equally culpable. Both were aware that they unleashed the same harms, and the simple fact is that neither had any (even partial) legally cognizable justification for doing so. Thus, Alexander and Ferzan's view has trouble with this case. The view seems to place no limit on how culpable one can be for even a minor offense provided only that one did it for bad enough reasons.[48]

Alexander and Ferzan's view also faces some internal tension. On the one hand, they acknowledge that motives don't matter when one's action is justified.[49] As long as the actor is aware of the applicable 'justifying reasons, it should not matter that the actor is not motivated by those reasons'.[50] Thus, they would say in EVIL

[46] *Id*. at 19. *See also id*. at 50.

[47] *See id*. at 27 (claiming that 'even very tiny risk impositions can be culpable if imposed for insufficient or misanthropic reasons').

[48] For similar reasons, Alexander and Ferzan's view also seems an awkward fit with the case of TED v.2. Ted's violation was caused by a huge amount of ill will towards Muslims. Still, his culpability for the fairly minor offense he committed is not as massive as would be suggested by the exceptionally criticizeable reasons for which he acted.

[49] Alexander & Ferzan, *supra* note 14 at 60–61.

[50] *Id*. at 61.

TROLLEY TURNER that Darryl is just as free from culpability as the analogous trolley turner with benevolent motivations. On the other hand, Alexander and Ferzan's theory also allows one's reasons (or motives) to affect culpability where the act is an unjustified violation.[51] Accordingly, in GOOD THIEF, BAD THIEF, they would allow that Barry's bad motives can make his theft more culpable than Mary's similar act done for more sympathetic reasons – although this sits uncomfortably with posited law.[52]

I don't claim that Alexander and Ferzan's view in GOOD THIEF BAD THIEF is *normatively* implausible. Indeed, I want my theory to leave room for this as a possible normative position (while also explaining as sensible the default irrelevance of motives in posited law). My main point instead is that it seems unstable to maintain both that motives do not impact the culpability calculus for justified actions of the sort seen in EVIL TROLLEY TURNER, but that motives do affect the culpability of unjustified acts as in GOOD THIEF, BAD THIEF.[53] This may not be a decisive objection, but I think there are benefits to seeking a unified account that treats motives as impacting the culpability calculus the same way in all cases. My view aims to do this.

C. The Epistemic Approach

An altogether different approach to understanding manifestation is *epistemic*. It focuses on the level of insufficient regard we would *infer* the actor possesses:

Epistemic approach: An action A manifests a level of insufficient regard *n* if and only if (and because) a rational, unbiased observer would infer from the relevant evidence that the defendant who did A possessed a level of insufficient regard equal to *n* when doing A.

What does 'the relevant evidence' mean here? It can't mean *all the facts*. Then the theory would say an action manifests as much insufficient regard as a *fully informed* observer would infer the actor possessed. But the fully informed observer would of course know

[51] For instance, they agree that their theory commits them to the claim that 'because the purpose of Frankie's trip is illicit [when she carefully drives off to kill Johnny], all risks she adverts to are themselves illicit' and render her culpable even for her careful driving. *Id*. at 50. Frankie's culpability here is due just to her bad reasons.

[52] *See supra* note 42–43.

[53] Compare Alexander and Ferzan *supra note* 14 at 50 with *id*. at 59–60.

what level of insufficient regard the actor actually possessed. Thus, she would recognize that negligent Charlie had just as little regard for others as his murdering counterpart Dennis. So the theory gets KILL YOUR UNCLE wrong. Instead, 'the relevant evidence' more plausibly means *the available evidence* – i.e. the reasonably obtainable evidence that can be introduced at trial. Typically, the available evidence would not allow us to infer that Charlie had any mens rea but negligence. After all, were Charlie called to answer for his uncle's death, he likely would disavow any intent to kill.

This view can be strengthened further. Gideon Yaffe is developing a more sophisticated version of the epistemic approach, which is similar in spirit to the theory I develop below. The view is still under construction, so I hesitate to discuss it in detail. But one fixed feature of his view deserves mention. He postulates a *principle of lenity*, which 'requires us to determine what the defendant's conduct says about his [level of insufficient regard] under the assumption that he is as little different from the law-abiding citizen as possible, given his behavior'.[54] This principle, *ceteris paribus*, 'bars us from being neutral between the hypothesis that the defendant is bad and the hypothesis that he is *very* bad'.[55] Thus, there is a *rebuttable presumption* that Charlie only acted with the insufficient regard of negligence, not the level of murder.

However, I doubt that either of these refinements – without more – is sufficient to render the epistemic approach defensible. In particular, one wonders about the case where the available evidence conclusively shows that Charlie in fact had the intention to kill, and thus a correspondingly higher level of insufficient regard. The prosecution might call a psychologist to testify, based on in-depth interviews (or brain scans), that Charlie actually intended to kill. More simply, perhaps Charlie is now overwhelmed by guilt and just *admits* on the stand that he did in fact intend to kill his uncle. Such scenarios are at least possible. In such a case, a reasonable observer would justifiably infer from the available evidence – even given the principle of lenity – that Charlie was not just negligent, but had the intent to kill. Accordingly, the epistemic approach, even with these refinements, entails that Charlie's action manifests the level of

[54] GIDEON YAFFE, AGE OF CULPABILITY: CHILDREN AND THE NATURE OF CRIMINAL RESPONSIBILITY, chap. 3, at 30 (forthcoming 2017).

[55] *Id.*

insufficient regard associated with murder – not just negligent homicide.

The difficulty here is that *the available evidence* is highly changeable. Often, the available evidence will only allow us to infer that defendants like Charlie are negligent. But sometimes the evidence would clearly show them to be willing to behave in far worse ways than they actually did. The epistemic approach does not fully preclude fixing their culpability by reference to the far worse acts we know their low levels of regard would make them *willing* to do.[56] That, however, conflicts with fundamental commitments of the criminal law.[57]

The epistemic approach likewise has trouble capturing the rule that motives generally don't matter to culpability. Suppose in EVIL TROLLEY TURNER, we're quite sure Darryl's actual motivation for turning the trolley was just to kill Vicky. Perhaps he simply admits it on the stand. The epistemic approach then entails that Darryl would not get the benefit of a necessity defense. But this conflicts with a widespread view in the law. Similarly for GOOD THIEF, BAD THIEF. Suppose Barry admits he stole the money just because he was bored and wanted to gamble. The epistemic approach entails that Barry is more culpable than others who steal for more sensible reasons. But, again, that is not the law's view. It does not take worse motives to be the basis for convicting one of a worse offense. Moreover, even if one thinks the law's view here is mistaken,[58] our theory should at least be *able* to capture such a widespread aspect of the law.

Of course, it is always open to defenders of the epistemic approach to say that motives generally don't matter only because of practical reasons – i.e. only because it is usually hard to identify and evaluate defendants' actual motives for breaking the law. Nonetheless, I think this pragmatic explanation is unsatisfying. It does not capture the full strength of the principles to which the criminal law is committed. After all, in some cases our practical limitations will be alleviated – i.e. where we do know the defendant's motives and have

[56] For similar reasons, one might worry that the epistemic approach would sometimes allow one to be criminally culpable merely for one's attitudes even if they have not yet resulted in any action.

[57] *See supra* notes 1–2.

[58] However, the law's view has much to recommend it. It seems more charitable to defendants that a) they be given the benefit of defenses that exist on the facts as they believe them to be, and b) for the law to ignore particularly bad motives (like Barry's) and treat all offenders the same as if they broke the law for less egregious reasons.

no trouble evaluating them. And in such cases, the epistemic approach gives no reason not to punish Darryl and Barry to the full and much harsher level that their bad motives might seem to call for. But the law does not change its view on this point when our practical limitations are temporarily alleviated. Even when we know with certainty what Darryl's and Barry's motives are and how seriously deficient their regard for others is, the law does not depart from its stance that Darryl gets the defense and Barry is not to be convicted of a more serious offense.

Accordingly, I remain skeptical about the prospects for the epistemic approach – though I have not decisively refuted it. More work might yield better answers.[59] But we at least have reason to continue looking for alternative ways to understand manifesting insufficient regard.

III. TOWARD A BETTER THEORY OF CRIMINAL CULPABILITY

This section aims to formulate a better insufficient regard theory by providing the sought-after account of manifestation. Thus far, several partly overlapping data-points have emerged. These should preferably be given a *principled explanation*, which applies even in isolated cases where our practical limitations are alleviated. Two are fairly basic:

(i) *Act Requirement:* Only voluntary conduct can be criminally culpable. Merely possessing bad attitudes that one does not act on cannot be criminally culpable.

(ii) *No punishment without prohibited conduct*: A legal system will not attribute culpability to you unless you've violated one of that system's prohibitions (though in principle these prohibitions could be norma-

[59] Perhaps we could adopt certain presumptions – based on principles of political morality – to constrain the body of evidence from which we may draw inferences about the defendant's level of insufficient regard. If we can block the evidence suggesting that Charlie actually possessed a worse mental state than negligence – e.g. the psychiatric testimony or brain scans that reveal his intention to kill his uncle – perhaps the observer could be blocked from inferring that Charlie possessed more insufficient regard than the amount seen in negligence. However, I'm skeptical about this strategy. What guarantee is there that the problems won't just re-emerge in isolated cases where the available evidence clearly rebuts the presumptions? Moreover, I think it will be hard to specify ex ante all the forms of evidence from which we might infer that the defendant was willing to act in worse ways than she actually did. It seems especially difficult to rule out such evidence stemming from *admissions* by the defendant that she had worse mental states or was willing to do worse things than her conduct at first revealed.

tively mistaken). One is not criminally culpable if one complies with the law – even if it is for very bad reasons.

Moreover, we have seen several data points relating to the idea that one should not be punished for bad attitudes, or a willingness to offend, unless it is manifested in action:

(iii) *Punishment only for actual conduct, not for willingness to behave in worse ways:* The criminal law should not attribute greater culpability merely because one would be willing to act in worse ways than what one actually did. Examples include TED THE WOULD-BE TERRORIST v.1 and v.2, as well as KILL YOUR UNCLE.

(iv) *The Concurrence Requirement*: To be guilty of a crime, the defendant's mens rea must concur with (i.e. actuate) the actus reus of the crime.[60] This means one cannot be punished merely on the basis of mental states acquired *after* doing the actus reus. Rather, some causal-explanatory nexus is needed between mens rea and actus reus.

However, (iv) must also be tempered by the following data point, which posed a problem for the pure causal approach to manifestation:

(v) *The amount of 'ill will … an action manifests is not the same as the amount … that exists and is being acted on'*.[61] Thus, one's act may not manifest the full amount of insufficient regard or animosity that actually impelled one to act – as when very bad attitudes cause one to act in ways that are only slightly criticizeable. An illustration is TRIVIAL WRONGS.

Finally, our theory must explain the complexities of the role of motives in the criminal law:

(vi) *Motives generally don't matter*: Substantive criminal law doctrine usually is not concerned with one's motives for acting. First, as (i) suggests, one is not criminally culpable for acting lawfully even for bad reasons. An example is EVIL TROLLEY TURNER. Second, substantive criminal law doctrine usually is not concerned with one's reasons for violating the law. An illustration is GOOD THIEF, BAD THIEF. But, third, there are exceptions to the default irrelevance of motives – as with (a) crimes of

[60] *See supra* note 3.

[61] Arpaly and Schroeder, *supra* note 18 at 188.

which bad motives are an element and (b) the role of motive in sentencing – which must also be explained.

I will defend a theory that captures these data-points. It has two components: (a) a necessary condition on culpability, i.e. a Manifestation Requirement, and (b) an account of *how much* insufficient regard is manifested. Under (a), culpability requires that insufficient regard is a cause of one's act. Under (b), how much the act manifests is largely independent of how much one possesses. I explain each component in turn. But first, the notion of culpability must be clarified.

A. Two Notions of Criminal Culpability

I have largely ignored an important distinction to this point. We must distinguish between the amount of culpability a jurisdiction actually attributes and the amount it ideally should. Call the jurisdiction-specific notion *posited culpability*. Suppose a jurisdiction recognizes no distinction between starting a fire in a building while knowing it is occupied and doing so without any such knowledge. Both types of arsonist are convicted of the same crime and subject to the same range of sentencing options. There is a sense in which this jurisdiction attributes the same amount of culpability for both forms of arson. They thus have the same amount of posited culpability.

However, there is also a sense in which this is normatively implausible. The same amount of culpability plausibly should not be attributed to these two forms of arson. The one arsonist disrespects the value of human life far more than the other. This suggests a *normative notion of criminal culpability*, which is what the law ideally should track. Normative criminal culpability is especially important for critical purposes. We can meaningfully ask how much normative culpability an action possesses even if it has not yet been criminalized. Or we might ask whether the posited culpability a jurisdiction attributes for an offense matches its normative culpability.

The insufficient regard theory can easily capture this distinction. It claims that an action is criminally culpable to the extent it manifests insufficient regard for legally protected interests or values – i.e. the legally recognized reasons that bear on whether to do the act.[62] We

[62] This idea of legally recognized reasons is also used by Gideon Yaffe. *See* Yaffe, *supra* notes 21 and 54.

can capture this distinction, then, by distinguishing (i) the interests, values or reasons that the law *actually* recognizes as bearing on whether to do the act from (ii) the ones that it *should* recognize. We can identify the interests, values or reasons in the former group by looking, e.g., at how harshly the jurisdiction punishes different types of conduct. The ones in the second group (the ones the law should recognize) are discernable via normative theorizing and policy assessment.

This distinction helps account for data-point (ii). For an action to possess *posited culpability*, it clearly must violate an existing criminal prohibition. After all, no well-functioning legal system that respects the principle of legality would permit punishment for an action that does not violate an existing prohibition. But this is not required for an action to be *normatively* criminally culpable. The law in a given jurisdiction might fail to protect certain interests or values (i.e. recognize certain reasons) that it ought to. One can manifest insufficient regard for interests or values the law *should* protect without violating an existing prohibition. Therefore, data-point (ii), only is a necessary condition on posited culpability, not normative culpability.

While the theory I offer below is primarily designed to capture posited culpability, the same framework also accounts for normative culpability. To do so, we simply need to ask how much insufficient regard an action manifests not for the interests and values that the law *actually* seeks to protect (i.e. the reasons it actually recognizes), but rather for the ones it *should*.

B. The Necessary Condition: Inadequate Repulsion from Criminality Must be a Cause of Your Act

Focusing on posited culpability for now, we reach the core question. On the present view, the criminal law's basic demand is not to avoid *possessing* insufficient regard, but to avoid *manifesting* it in conduct. What does this mean? The causal approach offers promising start: To manifest insufficient regard, an action must be caused by it. But as seen above, the pure causal approach fails because our actions do not necessarily manifest the full amount of insufficient regard that caused them. Thus, causation by insufficient regard seems to be only a

necessary condition for culpability. The *amount* manifested is a separate issue, discussed below.

This idea of causation by insufficient regard requires clarification. Importantly, it is not the same as being caused by animus. Actions can be culpable without such a cause. Instead, we can profitably understand this idea as failing to be sufficiently *repelled* by criminality. This happens when the legally recognized reasons against the action (i.e. the protected interests or values it threatens) fail to motivate one to avoid doing it. Causation by lack of contrary motivation[63] thus lies at the heart of an action's being caused by insufficient regard.

This does not mean the law requires one to be motivated *by* the legally recognized reasons to comply. Its demand is only that one have some motivations or other that prevent one from breaking the law. A convenient way to encapsulate this is to say that the law requires us to have some kind of *repulsion mechanism* whose job it is to repel us from criminality and to see to it that we are motivated to always act in legally justifiable ways. By 'legally justifiable action', I mean an action such that, under the facts *as you believe them to be*, the legally recognized reasons in its favor outweigh those counting against it. Thus, the repulsion mechanism is supposed to be responsive to the balance of available legal reasons that bear on your actions, and its task is to ensure that you always are motivated (in some way or other) to behave in ways that are supported by the balance of legal reasons. (I include the limitation to the facts as you believe them to be because criminal culpability is supposed to be largely a subjective notion, not just an assessment of whether one lives up to an objective standard – as in tort law.)

Crucially, the law is indifferent to the content of one's repulsion mechanism. Any motivations can in principle do the job.[64] Regardless of whether you are motivated to comply with the law for self-interested reasons, out of respect for law, to be seen as respectable, or for a hodge-podge of different reasons, the criminal law has no complaint against you. Its basic demand is only that some motivational failsafe kicks into repel you from criminality when necessary.

[63] Arpaly and Schroeder convincingly defend this sort of causation as scientifically respectable. *See supra* note 18.

[64] And when it fails to keep you within the bounds of legally justifiable conduct, it has the secondary job of motivating you to commit only the least bad violation possible (i.e. the one least disfavored by the legal reasons).

This begins to capture the law's default indifference to motives. To summarize:

Insufficient regard as repulsion failure: An action, A, that violates a legal prohibition[65] (i.e. satisfies the elements of a crime and is not otherwise justified) *manifests insufficient regard* for legally protected interests or values (i.e. the legally recognized reasons) only if the actor's repulsion mechanism is part of what *caused* A.

What, then, is involved in causation by repulsion failure? Two requirements must be met:

Causation by repulsion failure: Your action A was caused in part by a failure of your repulsion mechanism if:

(a) Your repulsion mechanism was *actually called upon to do some work* under the circumstances – i.e. called on to provide motivation against doing A, since it would be an unjustifiable violation of the law under the facts as you believe them to be – and
(b) Your repulsion mechanism failed to do the job it was called on to do (i.e. provide enough motivation to behave as required), such that you actually went on to perform A.

Only when the failure of your repulsion mechanism helps *cause* your conduct in this sense[66] is it *manifested* in that conduct. After all, even if the safeguard mechanism whose job it is to keep you within the bounds of legally justifiable conduct is faulty, this failure won't be *manifested* until the mechanism is actually called on to do its job, but doesn't. Before that, the failure of the mechanism is just a latent defect that has not yet been manifested.

What, then, does it mean for one's repulsion mechanism to be *called on* to do work in motivating one not to act unjustifiably? For posited culpability, it is fixed by the prohibitions in the jurisdiction.

[65] To get an account of normative culpability, just replace this with 'a prohibition the law *should* recognize'.

[66] On this account, the repulsion failure is usually going to be a but-for cause of the action. However, one could imagine Frankfurt-style over-determination cases where the repulsion failure (i.e. one's insufficient regard) *actually* led one to do the act, but where the repulsion failure was not a but-for cause. Perhaps other forces would have intervened to get one to do the action if the repulsion mechanism hadn't failed. Thus, I think repulsion failure strictly speaking needn't be a but-for cause of the action. Rather, it just has to be an *actual* cause of the act – on whatever the best account of causation turns out to be. (Thanks to an anonymous reviewer for pressing me on this point.)

In passing a criminal code, the legislature takes a stand on the normative question of when citizens' repulsion mechanisms should be called on. Of course, the legislature may get this *wrong*. Its laws might call on the repulsion mechanism to motivate one in ways that it should not. Criticizing the law in this way moves us into the realm of normative culpability.

Indeed, there are intuitive limits on when it would be fair to expect the repulsion mechanism to do work. It would be a blatant mistake to call on us to be motivated to avoid that which is impossible to resist. If someone pushes you off a bridge, your repulsion mechanism cannot be called on to motivate you not to fall. Similarly, it seems unfair for your repulsion mechanism to be called on to do work when you are neither engaged in any (relevant) conduct nor have any duty to act. When you are neither acting nor failing to discharge a duty, you are having no relevant effect on anyone else. In such cases, it would be odd and unfair for your repulsion mechanism to be called on to do any work. If that's right, the repulsion mechanism should not be called on to avoid bad attitudes that are in no way revealed in action. As a result, one cannot be *normatively* culpable merely for possessing bad attitudes. (Section IV further argues for this.)

Putting all this together, we get the following necessary condition on criminal culpability:

Manifestation requirement: The conduct of a competent, practically rational actor[67] that violates a legal prohibition[68] is criminally culpable *only if* a failure in her repulsion mechanism is part of the *cause* of that conduct (in the above sense).

One might question whether all of criminal law really involves such a requirement. What about crimes that require specific bad purposes or desires? Do they also involve only the failure to be repelled by criminality? Treason, we saw, requires the overt *purpose* of aiding the enemy; hate crimes require the *desire* to harm those with certain

[67] This claim is meant to apply only where the actor is at least minimally competent or practically rational. One who doesn't weigh the reasons incorrectly but violates a statutory prohibition only due to, say, a psychotic episode should not be deemed to have manifested insufficient regard. Thus, excused offenses are excluded from the present claim.

[68] Again, for normative culpability, just replace this with 'a prohibition the law *should* recognize'.

racial, gender or other protected characteristics – not just failure to be repelled by this result.[69] So don't these crimes involve an overt attraction to evil, not insufficient repulsion therefrom?

The main answer is that to be affirmatively attracted to a bad state of affairs is to *overvalue* the reasons in favor of bringing it about, which is equivalent to proportionally undervaluing the reasons against doing so. In my view, culpability depends on whether one attaches weight to the relevant reasons in the right *proportion* – not the absolute amount of weight one gives them.[70] Thus, when one overvalues the reasons in favor of a crime and doesn't attach correspondingly greater weight to the reasons against, one can be described as *undervaluing the reasons against the crime in the proportional sense* (i.e. compared to the weight attached to the reasons in favor of the act). Hence, treason and hate crimes involve *proportionally undervaluing* the reasons not to act with attitudes like racial animus or a desire to aid the enemy – a form of insufficient regard. (This worry also admits of other answers.[71])

Introducing this Manifestation Requirement already goes quite a ways to capturing the data from above. It explains data-points (i), and parts of (iv) and (vi). First, it captures the voluntary act requirement in (i) and the related idea that one cannot be criminally culpable merely for possessing bad attitudes. After all, you do not count as *manifesting* insufficient regard when you merely possess bad attitudes that don't produce any action. This, in turn, is because the necessary condition above is not satisfied. Under existing law (and any plausible set of normative assumptions), the repulsion mechanism would not be called on to do any work in such cases. It is only called on to ensure that you don't act in ways that cross the line into unjustifiable conduct (i.e. that are not supported by the balance of legal reasons that exist given the facts as you believe them to be). When you merely possess a bad attitude and don't act on it, there is nothing for

[69] *See supra* notes 7–9.

[70] *See* Sarch, *supra* note 17. I doubt the notion of the absolute weight given to a reason is even intelligible.

[71] In addition, the basic duty not to manifest insufficient regard can give rise to subsidiary duties – perhaps to prevent oneself from allowing certain proscribed attitudes from influencing one's conduct (i.e. to be so repelled by them that one keeps them in check). In this way, affirmative attractions to bad states of affairs can be culpability aggravators, which come *in addition* to the culpability of direct failures to be repulsed by bad states of affairs in their own right. *Cf. id.* (especially section 3).

your repulsion mechanism to kick into correct. So no failure of this mechanism is manifested.[72]

Second, the necessary condition helps explain some aspects of the concurrence requirement in (iv). It shows why you cannot be criminally culpable if you form the mens rea of the crime *after* doing the actus reus. There, too, no repulsion failure *actually helped cause your conduct*. For a failure of your repulsion mechanism to help cause your conduct, you at least had to possess the mens rea of the crime either *before or during* your performance of the actus reus.

Finally, the necessary condition helps explain part of (vi) – i.e. why motives don't matter to whether one possesses a defense. (Other explanations are available too, as noted below.) In EVIL TROLLEY TURNER, Darryl turns the trolley not in order to save the five but to kill the one on the other track. Why might this actor – despite clearly having insufficient regard – still get the benefit of the justification available to him based on the facts he was aware of? After all, he wasn't *actually* motivated by the facts that justify his conduct. One answer is that, although he possessed insufficient regard, it wasn't *manifested* in his conduct. Plausibly, his repulsion mechanism was not called on to do any work in this case. Since his conduct (turning the trolley) was justified on the facts as he knew them to be, no *failure* of his repulsion mechanism was part of what actually caused his conduct. Even though his repulsion mechanism was clearly *faulty* at the

[72] One might worry that my theory *could* in principle allow punishing merely for bad attitudes. (Thanks to Gideon Yaffe for this worry.) If the legislature criminalizes a particular attitude (e.g. disliking the supreme leader), then our repulsion mechanisms may be called on to motivate us to prevent ourselves from developing this attitude. Failing to properly manage one's own mental states to block this attitude from taking root might thus be a repulsion failure.While it's true that my theory thus has the resources to explain what is happening in such cases, that is a feature not a bug. For it's also true that my theory explains what has gone *wrong* in such cases: The legislature has endorsed a view about what work the repulsion mechanism is called on to do *that is obviously unjustifiable on normative grounds*. There are a range of moral, policy and practical reasons that explain why, in any world similar to ours, we should never punish merely for bad attitudes. (Section IV also mounts a *principled* argument for this claim.) This explains why the repulsion mechanism should never be called on to provide motivation to block the development of bad attitudes alone (as opposed to motivating us to prevent them from *impacting* our conduct, which can be legitimately required). Thus, although my theory does not render punishment for mere attitudes conceptually impossible, it does explain why this is would be prohibited under any *minimally plausible set of normative assumptions* – which I go on to justify directly in Section IV.

time, the fault remained latent. So no failure of the mechanism helped produce his conduct.[73]

C. *How Much Insufficient Regard is Manifested?*

The theory is not yet complete. We still must answer the question of *how much* insufficient regard an act manifests to know how culpable it is. As we saw, an action's being caused by a given amount of insufficient regard is not sufficient for manifesting that amount. My answer to this question employs a version of Yaffe's principle of lenity.[74] But while Yaffe's is epistemic, and constrains what we may infer from the available evidence, my principle is non-epistemic:

Principle of lenity: D's action, A, only manifests *the least amount* of insufficient regard for legally protected interests or values (i.e. the least amount of error in weighing the legally recognized reasons) that is needed to explain why a rational and otherwise well-motivated person would do A (i.e. what D did under the relevant description[75]) in the circumstances as D believed them to be.

The motivation for this principle is that the state, given its superior power, should resolve any ambiguity in its punishment practices in favor of accused citizens. This is a principled way to resolve difficult normative questions: when in doubt, benefit the defendant. I've discussed other justifications for this principle,[76] and Section IV gives a further normative argument for it.

---

[73] Does my view thus mistakenly entail that we shouldn't punish 'factually impossible' attempts? Factual impossibility traditionally was not a defense to attempt charges. *See* Yaffe *supra* note 13 at 112. But perhaps my view entails the opposite? When Jane tries to kill Victor by sticking pins into a voodoo doll replica of him, my view might suggest she doesn't merit punishment because her repulsion mechanism wasn't called on to get her to avoid this act. If no repulsion failure caused her act, she isn't culpable. (Thanks to Kim Ferzan for this worry.)
However, my view can avoid this result. The question of what the repulsion mechanism is called on to do is a normative question for the legislature in each jurisdiction to resolve. Quite plausibly, though, the best normative view is that one's repulsion mechanism *is* called on to halt one's attempted misconduct even in cases of factual impossibility. On the facts as Jane believes them to be, putting pins in the doll will kill Victor. Thus, it's very plausible that the law *should* call on her repulsion mechanism to motivate her not to stick pins in the doll. Hence, Jane *would* be culpable for attempting to kill Victor in this case. In this way, my view can capture the result that factual impossibility is not a defense to an attempt charge.

[74] Yaffe, *supra* note 21.

[75] As discussed below, it is an important question how this action, A, is to be described. When determining the culpability of an action that already counts as a crime in the relevant jurisdiction – i.e. when calculating posited culpability – the salient description of the act is given by the elements of the applicable statute. So it will typically include not just the relevant body movements, but also any required mens rea or attendant circumstances. However, when there is no statute on point prohibiting the conduct – as when we're concerned with normative culpability – it will be up for debate what the most apt description is. Different contexts plausibly call for different descriptions.

[76] *See* Sarch, *supra* note 3 at 32 n.97; Sarch, *supra* note 17 at 23 n.74.

Of course, the devil is in the details. How do we determine what this 'least amount of insufficient regard' is? To start, I follow Yaffe in understanding insufficient regard for legally protected interests and values as *committing an error in how one weighs the legally recognized reasons* that bear on how to act.[77] Next, distinguish the insufficient regard *possessed* when acting from the amount the action *manifests*. Culpability is pegged to the latter, not the former. My view uses an idealized procedure to determine the magnitude of the error in weighing the relevant legal reasons that a given action manifests. This procedure compares the motivations (i.e. weighing of reasons) of (a) the rational, perfectly law-abiding citizen and (b) the citizen who does the crime but represents the *smallest possible departure* from the law-abiding person.

To flesh out the idea, let's use a model of the ideal case – i.e. the motivations of the perfectly law-abiding citizen ('PC'). PC weighs the reasons for and against any putatively criminal action, A, she could perform the *correct way*. That is, she assigns weight to the legally recognized reasons for or against A exactly in line with the law's view of how they should be weighed.

To be more precise, assume first that the facts as the defendant, D, in the actual case believes them to be *are true*. (After all, this is a subjective culpability inquiry we're engaged in.[78]) Given the facts as D believes them to be, suppose there is a set of legally recognized reasons *against* A.[79] Call this set of reasons **R**−. These reasons will usually involve the legislature's reasons for criminalizing acts of type A (i.e. the interests or values this conduct was criminalized to protect).[80] Suppose there is a certain amount of weight that, according to the law, *should* be attached to these considerations. Let 'X' denote this amount, the correct weight. For PC, R− = X.

For completeness, suppose there also are some legally recognized reasons *in favor* of A. Perhaps A would promote some legitimate interests or prevent unjustified harms to others. Call this set of reasons **R+**. Suppose there is a certain amount of weight that,

[77] Yaffe, *supra* note 21.

[78] *Cf.* Alexander, *supra* note 40 at 24 ('Moral permissibility turns on how things really are, not on what an actor believes … Culpability, on the other hand, is a matter of the actor's beliefs'.).

[79] To turn this into an account of *normative* culpability, rather than *posited* culpability, we would focus not on the reasons that the law actually recognizes, but rather the reasons that the law should recognize (whatever they are).

[80] Note that this is just a *hypothesis* about what the real reasons are. The law can get this wrong.

according to the law, an actor *should* attach to these considerations. Let 'Y' denote this amount, the correct weight. For PC, R+ = Y. Of course, because A is assumed to be unjustified, it must be the case that X > Y.

PC's key characteristic, then, is that she is motivated by all and only the legally recognized reasons bearing on A, and she weighs them correctly. Thus, she does not actually perform A. Now, the core question to determine how much culpability a criminal action manifests is this:

Core question: What is the *smallest possible departure* from the motivations of the perfectly law-abiding citizen (PC) that would get an otherwise well-motivated person to do the criminal action A under the circumstances as the actual defendant believes them to be?

That is, what is the *smallest amount of incorrectness* in the weights attached to the legally recognized reasons for and against A that is needed to get an otherwise well-motivated person to do the criminal action A under the facts as the actual defendant, D, believes them to be? This is going to be our measure of *how much* insufficient regard is manifested in D's performance of A.

The Core Question admits of a general answer, indicating how culpable D's actual action is.

Answer to core question: The smallest amount of incorrectness in the weights attached to the legally recognized reasons bearing on criminal action A that would lead an otherwise well-motivated person to do A given the facts as the actual defendant believes them to be equals *an amount just infinitesimally larger than X minus Y* (i.e. the correct weight to be attached to the legally recognized reasons against A, or R−, minus the correct weight to be attached to the legally recognized reasons in favor of A, or R+).

Why is that? Call the otherwise well-motivated citizen OC. This person represents the smallest departure from the motivations of PC that would be needed to get someone to do A under the circumstances as D believes them to be. To get someone to do A, more weight must be attached to R+ than to R−. This means that the *least possible amount* that OC's valuation of the weights of the reasons for and against A must be off the mark by in order to get her to do A is just *slightly more* than X − Y. If R+ = R−, the actor's valuation of the reasons bearing on A would be in equipoise and no action would result. So what's needed is an amount slightly greater than this.

But what is the proof? How do we *know* the least amount of incorrectness in weighing reasons that would be needed to get OC to do A is just slightly greater than X − Y? To see why, note that there are three ways OC might attach more weight to R+ than to R−, and thus do A:

(1) OC attaches the right amount of weight to R+ (i.e. Y) but undervalues the reasons in R− so the total weight of R− dips down from X to a point below Y.
(2) OC attaches the right amount of weight to R− (i.e. X), but overvalues the reasons in R+ so the total weight of R+ rises from Y to a point above X.[81]
(3) OC both undervalues R− somewhat (so it's below X) and overvalues R+ (so it's above Y), and these mistakes together are big enough to make it the case that R+ > R−.

In all three cases, the *least* amount of error in weighing reasons that's required to get R− to be lower than R+ is the same: It's just slightly more than the difference between the correct weights for these two sets of reasons, viz. X and Y, respectively. Think of it like the levers on a sound mixing board, with one lever for R+ and another for R−. Ideally, the R− level should be at X, and the R+ level should be at Y (where X > Y). The smallest total distance that the two levers would have to traverse so that R+ > R− is just infinitesimally greater than X − Y. If they in total traversed a distance exactly equal to X − Y, then the two levers would not cross; at most they could end up at the same place, so that R+ = R−. Thus, for the R+ lever to *pass* the R− lever, such that R+ > R−, the smallest distance the two levers must traverse – the smallest error in weighing reasons that's needed – is an amount just slightly greater than X − Y. Using this model, we can now characterize *how much* insufficient regard an unjustified action manifests:

Level of manifestation: The amount of insufficient regard manifested in defendant D's unjustified action (assuming D is at least minimally competent and practically rational) directly corresponds to the *least amount of error* in weighing reasons that is needed to get an otherwise well-motivated person (OC) to do A under the circum-

[81] Note this could happen either by overvaluing some legally recognized reasons that belong in R+ , or by erroneously including reasons in R+ which aren't legally recognized as counting in favor of A at all.

stances as D actually believed them to be. That is, action A *manifests* an amount of insufficient regard that is just slightly greater than X − Y (as defined above).

A manifests an amount of insufficient regard just slightly in excess of X − Y because that is the smallest amount of error in weighing reasons needed to transform PC into OC. This is a principled approach. The amount manifested is not just a matter of how offended or harmed the victim is. Nor is it a matter of how bad we can infer the actor's response to reasons is given the available evidence, as Yaffe frames his view.[82] This approach would be unreliable because the evidence might be incomplete or inaccurate. My approach side-steps such evidential problems.

By adding this second component to the theory, we can explain the remaining data-points: (iii), (iv), (v) and (vi). Data point (iii) was that the law should not attribute greater culpability merely because one was willing to act in worse ways than one actually did. Recall TED THE WOULD-BE TERRORIST v.2. The salient description of what he did is given by the violation he committed: defacing public buildings with racist slogans. Thus, his culpability corresponds to the least amount of insufficient regard for legally protected interests and values that it would take to get OC, an otherwise law-abiding citizen, to behave as Ted did. Suppose the perfectly law-abiding citizen, PC, would recognize no reasons in favor of this action (as with a great many crimes). Moreover, given the interests and rights of the victims, PC would attach five units of weight to the reasons against this action. Thus, the smallest amount of error in weighing reasons that could get OC to behave as Ted did is an amount that is slightly greater than five units. (The size of the units doesn't matter, so long as they're consistent across cases.) Granted, Ted wanted and was *willing* to behave in far worse ways. And Ted might have *actually* assigned, say, ten units to the reasons in favor of this act and just 0.1 units against. But this is neither here nor there. The full amount of insufficient regard he possessed and acted from is not manifested. It reflects only a character flaw of the sort the law ignores. In this way, my theory explains why he is not more culpable just because his bad attitudes make him willing to behave worse than he actually did.

[82] *See generally* Yaffe, *supra* note 21 (discussing the 'inferences' we may draw about culpability from the defendant's conduct); *see also id*. at 9 (observing that '[i]nformation about the agent's psychological state at the time of action gives us information about the' factors that bear on culpability, and '[t]he principle of lenity mandates that we assign [culpability levels that are] as low as possible, consistent with the evidence').

For similar reasons, the theory explains the concurrence requirement in (iv). It states that a mens rea you possess when acting is appropriately connected to the actus reus only if the former *caused* (actuated) the latter. Recall KILL YOUR UNCLE, where Charlie negligently killed someone who turned out to be the person he intended to murder. Charlie is not guilty of murder, only negligent homicide. To account for this under the concurrence requirement, one might argue that Charlie's intent did not play the right kind of causal role in producing his action – though specifying what this causal role is appears difficult. But my theory offers a simpler explanation of this case. The amount of insufficient regard manifested in an action equals the least amount of error in weighing reasons that's needed to get OC to do the same conduct under the circumstances. The least amount of insufficient regard needed to get OC to perform Charlie's actual conduct – i.e. cause a death through carelessness – is the amount seen in negligence, not the far greater level associated with intentional killing. PC will appreciate and be motivated by the legally recognized reasons to pay adequate attention to the road while driving. Thus, the smallest error in attaching weight to legal reasons needed to get OC to behave as Charlie did is the amount involved in not attaching enough weight *to the reasons to pay attention to the road* – i.e. the amount in negligent unawareness of the risks he imposed.[83] So this is all the insufficient regard Charlie's actual conduct manifests. Even if he was willing to act worse than he did, his conduct does not *manifest* this fact about him.[84]

[83] A reviewer objects that my view does not explain Charlie's sort of inadvertent negligence, which does not involve the deliberate choice not to pay attention to the road. However, my view does not require a choice or the conscious weighing of reasons in order for an action to be culpable. Rather, all it requires is that the applicable legal reasons were accessible to one but failed to motivate one to abstain from the prohibited conduct. (Failing to pay attention to the road does not have to stem from a decision to attend to something else.) Nonetheless, on my view, negligent act A, done while being distracted in ways one ought not to be, can be culpable if A is caused in part by the failure to be sufficiently motivated to pay attention to the road. The insufficient regard A manifests lies in the failure to attach sufficient weight – whether consciously or unconsciously – to the legal reasons that exist to pay attention to the road and thus avoid A. Of course, this is just a sketch of how my account might explain inadvertent negligence. Much more must be said to adequately address this difficult topic.

[84] We could also get this result from the Manifestation Requirement. The only repulsion failure that was causally active in producing Charlie's conduct was the failure to be motivated to pay sufficient attention to the road. Had Charlie had gone on to commence an intentional killing, his repulsion mechanism would have been called on to motivate him to cease it. But he never got to that point. So this more egregious repulsion failure wasn't any cause of his actual conduct. In fact, his repulsion mechanism was only called on to motivate him to pay sufficient attention to the road, which it failed to do. Thus, only this smaller repulsion failure was actually causally active.

My theory also captures the additional wrinkle in (v). This was Arpaly and Schroeder's point that one is not always culpable to the full extent of the bad attitudes 'that existed and were acted on' – even if these attitudes *were* causally active in producing the act in question. Recall TRIVIAL WRONGS. There, the actual cause of Jill's decision to use the software to skim small amounts of money from her victims' accounts was bottomless ill-will. Nonetheless, the insufficient regard Jill's act *manifests* seemed rather small. The reason, on my view, is that the minimum amount of error in weighing the relevant legal reasons needed to get OC to do Jill's action under the circumstances as she believed them to be is *far less* than the amount she actually possessed and acted on. Thus, the full extent of her bottomless ill-will was not manifested in what she did.

The same reasoning also allows us to answer a common objection to the Model Penal Code's implicit claim that a knowledge crime is always at least *ceteris paribus* more culpable than the analogous recklessness crime.[85] Suppose D1 disregards a substantial and unjustified *risk* of death (say, a 30% chance) but is thoroughly callous about it. He cares not at all for his victims. By contrast, D2's situation is exactly the same as D1's except that D2 *knowingly* causes a death (i.e. is aware of a practical certainty that it will result), but has great regret and hesitation about doing so. Suppose these two cases are otherwise identical: Neither D1 nor D2 has any justification for their acts. Isn't D1 worse than D2? Why can D2 be convicted of a more serious offense than D1?

Again, the answer is that while D1 may be a worse person, this fact is not manifested in his conduct. D1's act does not manifest the full amount of insufficient regard he *possessed*. Rather, it manifests the least amount of insufficient regard – the smallest error in weighing reasons – needed to get OC to do this reckless act under the circumstances as D1 believed them to be (i.e. impose a perceived 30% chance of death without justification). But this is *less* than the amount of insufficient regard that D2's act manifests – i.e. the least

[85] *See* Kenneth Simons, *Punishment and Blame for Culpable Indifference*, 52 INQUIRY 143, 146 (2015).

amount needed to get OC to *knowingly* cause death (or recognize she is making it practically certain) without justification.[86]

Finally, my theory can account for the complex role of motives in the criminal law. However, to do so, we need to attend to an issue I've glossed over thus far: How are we to describe *what the defendant did* when assessing culpability? Where the relevant description does not mention the defendant's specific motives for breaking the law (only the actus reus he did and the mens rea that caused him to do it), my theory straightforwardly explains why one's subjective motives or aims do not impact the amount of insufficient regard manifested in one's action. After all, on my view, it would equal the minimum amount of insufficient regard (i.e. the smallest amount of error in attaching weight to reasons) needed to get OC to do the same conduct so described.

This gives a straightforward to way to capture cases that illustrate the default rule. In EVIL TROLLEY TURNER, if what Darryl did is described without mentioning motives – i.e. turning the trolley while aware that this will save five lives – it would not take any insufficient regard to get an otherwise well-motivated person, OC, to behave the same way. After all, the act is justified. Accordingly, even though Darryl was actually motivated by bad attitudes, the amount of insufficient regard his conduct manifests, thus described, is zero. Hence, my view is perfectly able to explain the law's position that the defendant gets the benefit of the defenses that exist on the facts he is aware of – even if these justifying circumstances did not actually motivate him.

GOOD THIEF, BAD THIEF can be handled the same way. If we use a motive-free description, Mary and Barry did the same thing: stealing something of value while aware that it did not belong to them. And

[86] Note one limitation. If there were facts about D2 that, for legal purposes, *partially* but not fully justified her conduct (e.g. if causing the death would prevent her friend from being injured), though these facts are not present in D1's case (i.e. D1 has no partial justifiers), then my view would not entail that D2 is more culpable than D1. All else is not equal between these two cases. It might well take less insufficient regard to get someone like D1 to recklessly cause death without any justification than it does to get someone like D2 to knowingly cause a death that is very nearly but not quite justified. (Thanks to an anonymous reviewer for pressing me on this point.)

Of course, the law of actual jurisdictions may not recognize all the partial justifiers it *should*. For example, Mary in GOOD THIEF, BAD THIEF should perhaps have a partial justifier that Barry lacks. But this is at most a problem for the jurisdiction in question, not my view. Were the law changed to give Mary a partial justifier, e.g. so she's guilty of a lesser offense, then my view could capture that result as well. My view makes posited criminal culpability a function of whatever the legally recognized reasons are. (And to determine normative culpability, we would need to take a position on what reasons the law *should* recognize – a question I cannot settle here.)

neither was aware of facts that amounted to a defense. (In the circumstances as they believed them to be, the theft was not necessary to prevent imminent bodily injury or death, etc.) Thus, it would take the same amount of insufficient regard to get OC to do what they both did. Thus, Barry (who stole for bad reasons) would not merit greater condemnation than Mary (who stole for not quite as bad, but still insufficient reasons).

Of course, some might reject the standard legal view on these issues. One might think it is normatively mistaken to ignore motives in such cases. Isn't Darryl worse than someone who turns the trolley to save the five? Isn't Barry worse than Mary, since she was motivated at least somewhat by altruistic considerations? It is a virtue of my account that it also can capture this view – if that's what one prefers. To do so, we simply need to adopt a thicker, motive-laden description of what was done. We get a harsher result in EVIL TROLLEY TURNER by describing what Darryl did as *turning the trolley for the sole purpose of killing one person and not at all in order to save the five*.[87] The least amount of insufficient regard needed to get OC to do this more richly described action would be quite high. So it would be very culpable. Likewise, it plausibly would take more insufficient regard to get OC to do the act of *stealing to afford gambling in Vegas* than *stealing to afford a fancier school for one's child*. Thus, Barry's act would be more culpable than Mary's. In this way, my theory can capture either intuition one has about these cases, which attests to its explanatory power.

Ultimately, it will be up to the legislature to decide what description to use. To benefit defendants in such cases and not attribute greater culpability to actors like Darryl and Barry, the legislature should define the applicable crimes in a motive-free way. By contrast, to prevent badly-motivated actors like Darryl from having a defense, and to impose harsher penalties on badly-motivated actors like Barry, the legislature should adopt a motive-laden description. Indeed, I think this is precisely what occurs for crimes like treason, kidnapping and hate crimes, where a specific bad purpose or motive is included as an element. For these crimes, the legislature decided that some kinds of bad aims or motives are especially worthy of condemnation (perhaps because there is a special duty to avoid being

[87] This is Tados's view about the case. *See supra*, note 41.

motivated by, e.g., racial animus, disloyalty to country or the desire to terrorize victims). Thus, my theory also explains what's going on with exceptions to the default rule that motives don't matter.

Still, this is only enough to account for posited culpability. What about normative culpability – that which the law *should* attribute? This notion is important for critiquing existing law and deciding whether to adopt new offenses or defenses. There, theorists must take a stand on how to describe the defendant's conduct and can't simply defer to the legislature's view.

I can't fully defend it here, but my view is that a coarse-grained, motive-free description will generally be more appropriate for substantive criminal law doctrine (i.e. the rules applicable at the guilt-stage), though I'm open to the idea that a more fine-grained, motive-laden description may be more appropriate at the sentencing stage, after the defendant has been convicted. Many get accused of crimes, but before a conviction is reached, it is not yet clear if it's warranted for the court to spend time and effort in closely examining the most minute details of the case. Prior to conviction, the question is simply which course-grained box to place the defendant into – that is, which offense to convict him of, if any, or whether he qualifies for one of the narrowly defined affirmative defenses. For this question, the defendant's mens rea (or what he was aware of) will be relevant, but typically not his motives. By contrast, once a verdict has been reached regarding what course-grained category the defendant belongs in, we can be confident that there is a basis for more closely considering the details of the case. Thus, at sentencing, a fine-grained, motive-laden description begins to seem more appropriate.[88] (I also think that at least in ideal conditions, the motive-free description is to be preferred at the guilt-stage because this is more

[88] *Cf.* 18 U.S.C. § 3553(a)(1) (including 'the nature and circumstances of the offense and the history and characteristics of the defendant' as one of the factors to be considered in sentencing).

beneficial to defendants.[89] However, this normative argument is beyond the scope of this paper.)

Accordingly, my theory accounts well for the complex role of motives in the criminal law. Usually, they don't matter because legislatures (reasonably) take the course-grained, motive-free description to be the most suitable way to define the act types to be criminalized. But sometimes legislatures take the opposite view – as we see with hate crimes or treason. Moreover, once the defendant is convicted, the perquisites are met for taking a closer look at what the defendant did. So motives might matter to the more fine-grained assessments in sentencing. (Of course, motives may also be relevant to sentencing because culpability is not the only relevant sentencing factor.[90] A defendant's motives in doing the crime, and subsequent attitudes like remorse, might help show what sentence is needed for specific deterrence and rehabilitation. So this provides a supplementary, or perhaps an alternative, explanation of why motives affect sentencing.)

---

[89] I develop the argument elsewhere (*see* Sarch, *The Moral and Legal Contours of Willful Ignorance*, chap. 3 (manuscript on file with author)), but here is the idea. The motive-free and the motive-laden approaches come apart in two types of case. The motive-*free* description is better for defendants when they are aware of some at least partially-justifying reasons for acting, but actually are motivated by worse reasons. EVIL TROLLEY TURNER is an example. By contrast, the motive-*laden* approach is better for defendants when they are actually motivated by considerations that are *better* than the best legally recognized case in favor of their actions, given their beliefs. An example might be Mary's more sympathetic motives in GOOD THIEF, BAD THIEF. You might think Mary should get a break and be punished less than Barry, since her motives are better than his.

It's hard to know which kind of case is more numerous in an actual jurisdiction. But we can at least settle the question in ideal circumstances. The key point is this: *There are no cases of this second type* when the law is normatively ideal and recognizes all the reasons that it should that can help justify one's conduct. Assuming the law is ideal in this sense, the legally recognized reasons that support the defendant's conduct given her beliefs about the case will always be *the best possible normative case that can be provided in favor of her conduct, given her beliefs*. For example, if Mary's theft really is at least somewhat (if not fully) justified by her need to be able to afford a better school for her child, then this is a consideration that the law should recognize in some way – either as a mitigating defense, or as the basis for convicting Mary of a less serious theft offense than someone like Barry (who stole without any comparable need). If the law is ideal in this sense, Mary's punishment would be reduced because the theft was done in circumstances where it promoted a legitimate need. Thus, in such a system, she'd get no benefit from an approach that considers her actual motives. Since the law is assumed to be ideal, this would already be baked into the elements of the offense or give her a defense. (Granted, this would give rise to EVIL TROLLEY TURNER type cases, where one might be aware of facts that confer at least a partial defense but one is not motivated by those justifying facts. However, that only redounds to the benefit of defendants.)

Thus, at least if the law is ideal – i.e. recognizes all and only the justifying circumstances it should – there will be no cases where considering the defendant's actual motives will benefit him in the culpability assessment. By contrast, there will still be cases where defendants are benefitted by *ignoring* his actual motives – as in EVIL TROLLEY TURNER cases. Therefore, assuming the law is ideal, the approach that benefits defendants most is to describe 'what he did' in a motive-free way when assessing culpability. Thus, for normative criminal culpability, a motive-free description should ideally be used (at least for substantive doctrine applicable at the guilt-stage).

[90] *See, e.g.*, 18 U.S.C. § 3553 (listing factors to be considered in sentencing).

D. Objections

Two objections demand a response. First, doesn't my theory impute culpability not just on the basis of the defendant's actual conduct – even though I've said this is something we generally shouldn't do? Two points in response. First, I grant that my theory relies on an idealized inquiry to determine how much culpability an action manifests. But this inquiry will never be relevant *unless* the failure of the defendant's repulsion mechanism *actually* helped cause her criminal action. Thus, my theory does not impute culpability merely for counterfactual attitudes or conduct. Instead, culpability is only imputed for actual conduct, and it is just the *amount* of culpability imputed that gets calculated using an idealized procedure. Second, any remaining aspects of my theory that might seem 'counterfactual' in nature all redound to the benefit of the accused. On my view, the defendant will never be punished *more harshly* using the idealized calculation I advocate than her actual mental states merit (i.e. the level of insufficient regard she possessed at the time). The idealized calculation in my theory only helps defendants, but doesn't hurt them. This should make us more comfortable with this idealized calculation.[91]

A second objection is this. Perhaps you think it's not a problem to *overvalue* something that really is valuable – say, your children's interests. How can it be bad, one might wonder, to care *too much* about something good? David Shoemaker recently raised such an objection to theories like mine: I could 'take extremely seriously your interests in doing what you want with your property but still

[91] One might object to this on the basis of what I said in discussing recklessness vs. knowledge crimes. *See supra* note 85 and accompanying text. Doesn't my ideal calculation impute more culpability to D2, the regretful knowing killer, than the amount of insufficient regard she possesses at the time of acting? Suppose D2 weighs the reasons against the killing at +10 (which is the correct amount) and the reasons in favor at +11 (which is well above the correct amount). It might seem that D2 possesses a fairly small amount of insufficient regard. Doesn't my ideal calculation impute more culpability to her than the amount she possesses?

No, the idealized calculation still imputes less culpability than the amount of insufficient regard she actually possesses. Here's why. Suppose that the weight D2 should have attached to the reasons in favor of the act was only +1. She thus overvalued R+ by a factor 11. This means her repulsion mechanism was called on to create a correspondingly large increase in the weight attached to the reasons against. After all, that's what's needed to preserve the proportional relationship between the actual weight of reasons (i.e. 10 against versus 1 in favor). But the mechanism didn't do its job and this helped cause her criminal act. Thus, my theory says we impute the most lenient possible amount of culpability to her for her action – i.e. slightly more than $10 - 1 = 9$ units. Although D2 in fact dramatically undervalued the reasons against the act (i.e. she assigned +10 to the reasons against when she should have assigned 11 times as much), she is treated far more leniently. In fact only +9 units of culpability are imputed to her. Thus, on my idealized calculation, she is in fact held to be less culpable than the full amount of insufficient regard she actually possessed at the time.

steal from you because I weigh my [children's interests] as slightly more important than yours'.[92] Here I seem to have sufficient regard, but am still criminally culpable.

Granted, overvaluing something good (like one's children's interests) may not be bad in itself. But it *is* a problem when it leads you to attach too much weight to your kids' interests as a reason to do a criminal act *compared to* the weight you attach to the other legally recognized reasons against doing that act. In that case, you do manifest insufficient regard for legally protected interests. On my view, an important aspect of criminal culpability is the *proportional relationship* between the weights you attach to R+ and R−. After all, this is what determines if your repulsion mechanism has failed. If overvaluing your children's interests leads you to attach a *disproportionally* large amount of weight to R+ (the reasons in favor of A) *compared to* the weight you attach to R− (the reasons against A), then your repulsion mechanism should kick into get you to attach correspondingly greater weight to R−. If it doesn't, and you end up overvaluing R+ without increasing the weight attached to R− sufficiently to get you to avoid doing A, then your repulsion mechanism has failed. That is, you end up *proportionately undervaluing* R−, and that is just what having insufficient regard for the reasons not to do A consists in. When such a repulsion failure ends up causing a prohibited act, that is the essence of criminal culpability.[93][94]

[92] David Shoemaker, *Blame and Punishment*, in BLAME: ITS NATURE AND NORMS 110 (D. Justin Coates and Neal A. Tognazzini, eds., 2012).

[93] That's why overvaluing the reasons in favor of a crime can be aptly described as *insufficient regard*. It's insufficient regard for the reasons against the crime as *compared to* the weight you happen to attach to the reasons you see in favor of doing it. (Nonetheless, as I've argued elsewhere, an overt attraction to the reasons in favor of a crime where no such reasons exist can still be a source of *heightened* culpability. *See* Sarch, *supra* note 17.)

[94] One might also object that my view does not explain why strict liability crimes are legitimate. One could be convicted of a strict liability crime even when my view says one has no criminal culpability – e.g. if one has no insufficient regard and does a strict liability crime purely by accident. The least amount of insufficient regard needed to get an otherwise well-motivated citizen to commit such a violation is zero. I accept this implication. My view does not entail that those convicted of strict liability crimes are culpable – and it's a good thing too. This is precisely what makes strict liability crimes troubling. Thus, on my view, if such crimes can be justified at all, it will have to be for consequentialist reasons pursuant to some hybrid theory of punishment.

E. Concluding Remarks: The Law as a Simplified Analog of the Moral Landscape

A major attraction of my picture is that it preserves the continuity between criminal culpability and moral blameworthiness. On quality of will theories, an act is morally blameworthy to the extent it manifests insufficient regard for the moral reasons bearing on whether to do that act. These moral reasons can in principle be as detailed and case-specific as you like. By contrast, on my view, an act is criminally culpable (in the posited sense) to the extent it manifests insufficient regard for the applicable legally recognized reasons. Very likely, the legislature cannot (and should not) recognize all moral reasons bearing on how to act. Some might think one is less morally blameworthy for a theft if the motive for it is to help one's child, but the legislature may reasonably decline to recognize this as a consideration that justifies theft even partially. Thus, the landscape of legally recognized reasons that determine culpability is, and likely should be, more anemic than the richer landscape of moral reasons that affect blameworthiness. (The law may also recognize reasons that map onto no corresponding moral reason.) My view thus preserves a structural analogy between criminal culpability and moral blameworthiness, while also respecting the differences between the two. The former tracks one's responses to a different, thinner set of reasons than the latter.

IV. AN ARGUMENT FOR THE MANIFESTATION REQUIREMENT AND THE PRINCIPLE OF LENITY

Thus far, I've been mainly concerned with descriptive adequacy. I have not said much about why the law *should* adopt its current skeptical stance toward punishing mere bad attitudes. Even if the data-points my theory aims to capture reflect *posited* culpability in Anglo-American systems, why think this matches the correct attributions of *normative* culpability? To fill this gap, let me sketch an argument for why one's view of culpability should include a Manifestation Requirement and Principle of Lenity, and thus deliver results roughly in line with existing law.

My argument draws in part on practical considerations and our epistemic limitations, but it puts them to use in a principled way.

The argument is supposed to explain why one-off cases of punishing merely for bad attitudes, or for one's willingness to offend under counterfactual circumstances, would not be permitted when our practical limitations are only *temporarily* or *locally* alleviated. Specifically, I offer a contractualist argument that supports broad principles that would be accepted in any world with the same stable background conditions as ours.[95]

Suppose that in entering into civil society, we agree not to engage in certain kinds of conduct, and cede to the state the right punish such proscribed conduct. In negotiating, under the veil of ignorance, what rules will govern our lives in civil society, it would make sense for us to only make the *least burdensome concessions* we have to in order to obtain the benefits of the social contract. Now consider three possible criminal justice regimes. In the *maximalist regime*, punishments are doled out according to the full amount of insufficient regard (i.e. bad attitudes) one happens to possess, regardless of whether it is manifested in action. In the *minimalist regime*, one is only punished for insufficient regard that is manifested in conduct, and the amount manifested is determined in accordance with a principle of lenity as my theory suggests. Third, the *intermediary regime* includes a rudimentary manifestation-like requirement but no lenity principle. Thus, one's culpability equals the level of insufficient regard that actually *causes* one's actions, even if it's not fully manifested under the lenity principle.

The minimalist regime, I contend, offers contracting parties the best deal. To see why, consider some cases of unmanifested mental states:

(a) Charlie drives off intending to kill his uncle, and in his distracted state hits and kills a pedestrian who just happens to be his uncle.
(b) Alan burns down an unoccupied building for $5000, though we know for a fact he'd be willing to burn it down for the same sum of money even if he knew a person was inside.

Both the maximalist and intermediary regime would take Charlie to be guilty of murder and Alan to be guilty of a higher grade of arson

[95] Of course, the argument thus inherits all the familiar problems with contractualist arguments. *See, e.g.*, Philip Stratton-Lake, *Scanlon's Contractualism and the Redundancy Objection*, 63 ANALYSIS 70 (2003). Space does not allow me to answer them here. But permit me to adopt a contractualist framework to merely *illustrate* the line of thinking that would support the Principle of Lenity and the Manifestation Requirement on which my theory relies.

that is appropriate when the arsonist knows a person was in the building. Charlie and Alan not only possess levels of insufficient regard that are associated, respectively, with murder and the higher grade of arson; in addition, their conduct is also caused by their high levels insufficient regard, and so would count as manifested in the intermediary regime. By contrast, in the minimalist regime, Charlie is guilty only of negligent homicide, and Alan is guilty of the lower grade of arson used when the building is unoccupied.

My claim, then, is that rational actors negotiating behind the veil of ignorance would prefer the minimalist regime to either the maximalist or intermediary regime. That is because the minimalist regime involves far less burdensome concessions than the other regimes, while also providing nearly the same benefits in terms of harm prevention. Thus, the net expected benefit of the minimalist regime substantially outweighs that of the other two.

Why is this? Begin with the benefits of the competing regimes. Agreeing to be punishable for the full amount of animosity or callousness towards others that we possess or is merely involved in causing our actions, even when its full extent is not manifested, would provide scant benefits compared to agreeing to be punishable only for the insufficient regard our conduct manifests in my sense. After all, when others harbor hateful thoughts or disrespectful attitudes towards us, but don't act in ways that bring the full badness of these attitudes to light, the extra harm we suffer from the unmanifested badness of these attitudes is extremely limited – if there is any extra injury at all. When Charlie or Alan are merely willing to act in worse ways than they did, but don't have occasion to manifest this willingness, there is little or no extra injury from the mere willingness. That means that there is little or no additional harm to be *prevented* in cases of unmanifested mental states (the only cases where the regimes differ) by imposing the greater punishments of the maximalist or intermediary regimes. The overall level of harm to be prevented is roughly the same if we focus on either (a) defendants' actions plus the full amount of insufficient regard they possessed or acted on at the time, or (b) their actions plus the insufficient regard these actions *manifest* (understood according to the principle of lenity). Thus, there is little call, on protective grounds, for preferring the maximalist or intermediary regime to the

minimalist regime. Assuming the three regimes already provide adequate harm prevention (i.e. a sufficient baseline level of deterrence), there will be no reason to move beyond the minimalist regime to the maximalist or intermediary version. There is no significant added harm from unmanifested insufficient regard that this is needed to prevent.[96]

Now consider the concessions involved. Most importantly, to accept the maximalist or intermediary regimes, and thus agree to be punishable for the full level of hateful thoughts, uncaring attitudes or willingness to harm that happens to play some causal role in getting one to act, would entail enormous enforcement-related costs compared to the minimalist regime.[97] Enforcing the rules of the maximalist or intermediary regimes would be heavily invasive. It would require allowing the state to investigate and concern itself with evidence of whether one possesses or was even partly motivated by insufficiently respectful attitudes – including one's utterances and writings. This would undermine the substantial benefits we get from privacy and unencumbered speech. These include the benefits of getting to consider dubious hypotheticals and mull over a range of thoughts and plans (some perhaps quite bad) before deciding how to act – not to mention the feeling of safety we might get from having a private, protected mental space. Thus, giving the state the amount of control over us required to implement the maximalist or intermediary regimes would involve heavy costs, which the minimalist regime largely avoids.

[96] An anonymous reviewer points out that the maximalist or intermediary regime might have advantages in providing more general deterrence than the minimalist regime. However, in comparing these regimes, we should keep all else equal. Thus, we should assume that the regimes all include a baseline level of punishment that gives adequate general deterrence. (It would be unfair to assume that the one regime provides deficient deterrence.) My point, then, is that there is very little additional harm from unmanifested bad attitudes that we would need to combat by ratcheting up the level of deterrence. Thus, assuming all three regimes are on a par in offering adequate deterrence, there is little call on protective grounds to move from the minimalist regime to something more burdensome like the maximalist or intermediate regime.

Consider an analogy. Suppose we are considering enhancing punishments for crimes committed by people who have a disfavored attitude like a preference chocolate ice cream. Let's say there is no independent harm in preferring chocolate ice cream. Granted, we could get more general deterrence from increasing punishments for those who commit crimes with this disfavored attitude. But assuming the existing punishments – without any such enhancement – do not leave significant harms undeterred, then there is little protective reason to enhance punishments *further* by moving to a regime that imposes harsher penalties on those who prefer chocolate ice cream.

[97] A further cost of the maximalist or intermediary regime compared to the minimalist regime is, of course, that the former entail greater punishments should we ever find ourselves in the position of Charlie or Alan, acting criminally without fully manifesting all our insufficient regard. The greater harms imposed by the maximalist and intermediary regimes on contracting parties who break the law also affect the calculus.

Accordingly, the bargain offered by the maximalist or intermediary regimes – i.e. agreeing to be punishable also for any bad attitude that one possesses (in the maximalist case) or is any part of what causes one to act (in the intermediary case) in exchange for others agreeing to the same – just does not seem worth it. The bargain offered by the minimalist regime, by contrast, with its Manifestation Requirement and Principle of Lenity, seems a better deal. It offers greater net benefits. The bargain that seems to do the best job of providing the negotiating parties with tangible and valuable benefits, in exchange for minimally costly concessions, is to agree to be punishable only for the degree of insufficient regard that is *manifested* in conduct, determined according to the Principle of Lenity (i.e. construed in the light most favorable to the accused).

Thus, the negotiating parties likely would prefer the minimalist regime. It is doubtful that those negotiating behind the veil of ignorance would rationally agree to be punishable for the full extent of the insufficient regard they happen to possess or that merely helps cause their actions – as in the cases of Charlie and Alan. The benefits we, as potential victims, get from extracting such concessions from the Charlies and Alans of the world are not very great – certainly not substantial enough to justify the steep costs of punishing us for our unmanifested bad attitudes.

School of Law,
University of Surrey, Guildford, Surrey, GU2 7XH, UK
E-mail: a.sarch@surrey.ac.uk