University of Southern California Law School
Legal Studies Working Paper Series

Year 2015 *Paper* 176

Double Effect and the Criminal Law

Alexander F. Sarch*

*University Southern California, alexsarch@gmail.com

Double Effect and the Criminal Law

Alexander F. Sarch

Abstract

American criminal law is committed to some version of the doctrine of double effect ("DDE"). In this paper, I defend a new variant of the agent-centered rationale for a version of DDE that is of particular relevance to the criminal law. In particular, I argue for a non-absolute version of DDE that concerns the relative culpability of intending a bad or wrongful state of affairs as opposed to bringing it about merely knowingly. My aim is to identify a particular feature of the former in virtue of which it is pro tanto more culpable than the latter. Providing an agent-centered argument of this kind for a culpability version of DDE, I argue, is an especially attractive route to take for those who are interested in vindicating the way the criminal law actually encodes DDE.

DOUBLE EFFECT AND THE CRIMINAL LAW

Alexander Sarch*

Forthcoming in *Criminal Law and Philosophy*
[Please cite only from published version]

ABSTRACT: American criminal law is committed to some version of the doctrine of double effect ("DDE"). In this paper, I defend a new variant of the agent-centered rationale for a version of DDE that is of particular relevance to the criminal law. In particular, I argue for a non-absolute version of DDE that concerns the relative culpability of intending a bad or wrongful state of affairs as opposed to bringing it about merely knowingly. My aim is to identify a particular feature of the former in virtue of which it is pro tanto more culpable than the latter. Providing an agent-centered argument of this kind for a culpability version of DDE, I argue, is an especially attractive route to take for those who are interested in vindicating the way the criminal law actually encodes DDE.

American criminal law is committed to some version of the doctrine of double effect ("DDE"). That is, various criminal law doctrines embody the idea that it is worse to intentionally (or purposefully[1]) bring about a bad or wrongful state of affairs than to merely foresee (or act with the knowledge[2]) that one's conduct will bring about that state of affairs.[3]

Treason, for example, requires acting with *purpose* to aid the enemy, not merely with knowledge that one's conduct will have this result.[4] In *Haupt v. United States*, the father of a German saboteur operating in the U.S. during World War II was convicted of treason for "[s]heltering his son, assisting him in getting a job, and in acquiring an automobile, all…with knowledge of the son's mission."[5] On appeal, the father argued there was not "sufficient proof of adherence to the enemy" to support a conviction, "the acts of aid and comfort being natural acts

* I would like to thank Steve Finlay, Joe Horton, Andrei Marmor, Jake Ross, Mark Schroeder and especially Jon Quong for extremely helpful comments and conversations about earlier drafts of this paper. I am also grateful to an anonymous reviewer for this journal for valuable feedback.

[1] *See* Model Penal Code § 2.02(a)(i) (defining the mental state of purpose). I use "intentionally" and "purposefully" interchangeably per the legal convention. *See infra* note 3.

[2] *See* Model Penal Code § 2.02(2)(b)(defining the mental state of knowledge as awareness or practical certainty). I use "foreseen effect" interchangeably with "effect one knows one will bring about."

[3] Wayne LaFave, 1 SUBST. CRIM. L. § 5.2 (2d ed.) ("the modern approach is to define separately the mental states of knowledge and intent (sometimes referred to as purpose…)").

[4] *Id*. (footnote 9).

[5] 330 U.S. 631, 633 (1947).

of aid for defendant's own son."[6] The Supreme Court rejected his argument, and found instead that it was for the jury to decide whether the father's "intention was not to injure the United States but merely to aid his son 'as an individual," or whether his purpose really was "aiding the German Reich, [and] injuring the United States."[7] The Court held there was indeed sufficient evidence to support the jury's finding that the father acted with purpose to aid the enemy. The father had said things to the effect that "after the war he intended to return to Germany, that the United States was going to be defeated, that he would never permit his boy to join the American Army, [and] that he would kill his son before he would send him to fight Germany."[8] The Court thus upheld the guilty verdict, concluding that the son was in fact just a "chip off the old block."[9]

Is the law correct to follow DDE in cases like *Haupt* and assume there is a normative difference between purposefully bringing about a prohibited state of affairs and doing so only knowingly? What are the arguments for thinking so? As Dana Nelkin and Samuel Rickless note, two main rationales have been given for DDE.[10] One is victim-centered, and is based on the "Kantian idea that persons have a right not to be used as means without their consent."[11] Thus, intentional harm seems especially troubling when and because it merely *uses* others for one's own ends. The other rationale, by contrast, is agent-centered. It is based on "the idea that there is something especially morally problematic about aiming at evil."[12]

In this paper, I defend a new variant of the agent-centered rationale for a version of DDE that is of particular relevance to the criminal law. In particular, I will argue for a specific formulation of DDE that concerns the relative *culpability* of intending a bad or wrongful state of affairs as

[6] *Id*. at 641.
[7] *Id*.
[8] *Id*. at 642.
[9] *Id*.
[10] Dana Nelkin & Samuel Rickless, *The Relevance of Intention to Criminal Wrongdoing*, forthcoming in CRIM. L. & PHILOSOPHY, p. 12 (http://link.springer.com/article/10.1007/s11572-014-9343-0).
[11] *Id*. *See also* Warren Quinn, *Actions, Intentions, and Consequences: The Doctrine of Double Effect*, 18 PHIL. & PUBLIC AFFAIRS 334 (1989).
[12] Nelkin & Rickless, supra note 10. *See also* Thomas Nagel, THE VIEW FROM NOWHERE 181 (1986).

opposed to bringing it about merely knowingly. My aim is to identify a particular feature of the former in virtue of which it is *pro tanto* more culpable than the latter.

In section 1, I formulate the version of DDE I will be concerned with and explain its relevance to the criminal law. Section 1 also explains why agent-centered rationales are especially attractive to those interested in the criminal law. In section 2, I sketch the basic argument for my culpability version of DDE. Section 3 refines the argument, while section 4 responds to objections. Section 5 wraps up by indicating how progress might be made on the so-called closeness problem. Although this problem ultimately might prove so intractable that the criminal law should be reformed to reject DDE, this is a conclusion I think we should adopt only as a last resort. I don't think we have to fall back to this position yet, though, since the argument I develop here gives reason to be optimistic that the criminal law's actual uses of DDE can be placed on a more secure normative foundation.

1. A Version of DDE for the Criminal Law

In this paper, I will focus on a non-absolutist, culpability-based version of DDE, and my argument for it is agent-centered. In this section, I explain why this sort of argument for this version of DDE is of particular interest to those seeking a normative foundation for the way the criminal law actually embodies DDE.

1.1 Why care about culpability-versions of DDE?

Different versions of DDE have been formulated using different normative concepts. Sometimes it is construed as a claim about permissibility—*i.e*. a claim to the effect that for some harms, it is impermissible to bring them about intentionally, but permissible to bring them about as foreseen side-effects of one's conduct.[13] Sometimes DDE is formulated as a claim about

[13] William FitzPatrick, *The Doctrine of Double Effect: Intention and Permissibility*, PHILOSOPHY COMPASS 7/3 (2012).

difficulty of justification—*i.e*. that it is harder to justify intentional harms than the analogous merely foreseen harms.[14] Others have understood DDE in terms of culpability,[15] which is the approach to formulating DDE I adopt here. (These different formulations of DDE of course might be related if the normative concepts they employ are connected in the right way.)

There are at least three reasons for those interested in the defensibility of existing criminal law to focus on a culpability version of DDE. First, the criminal law is often said to embody a *culpability hierarchy*. Modern criminal law employs four main mental states one might act with—negligence, recklessness, knowledge and purpose—each supposedly more culpable than the last.[16] Ken Simons notes that this hierarchy is embodied in the influential Model Penal Code ("MPC"):

> The MPC views its four basic mental states or culpability terms as hierarchically ordered: all else being equal, purpose is more culpable than knowledge, which is more culpable than recklessness, which is more culpable than negligence. Indeed, the MPC explicitly provides that if a statute requires a mental state that is lower in the hierarchy, then an actor who possesses a higher mental state also satisfies that mental state requirement.[17]

If we are interested in whether this culpability hierarchy is defensible, we will want to know to what extent its claims about the relative culpability of these four mental states hold. The distinction between the top two mental states in the hierarchy—purpose and knowledge—directly corresponds to the distinction in DDE. Thus, securing a culpability version of DDE would shore up the normative credentials of the culpability hierarchy.

A second, related reason to care about a culpability version of DDE is that showing it to be defensible would help explain certain criminal law doctrines. Most importantly, it would explain why it is sensible for the criminal law to condition guilt for certain crimes on having the purpose

---

[14] Ralph Wedgwood, *Defending Double Effect*, 24 RATIO 384, 385-86 (2011); Neklin & Rickless, *supra* note 10 at 13.
[15] MICHAEL MOORE, CAUSATION AND RESPONSIBILITY 48 (2009).
[16] Model Penal Code § 2.02(a)-(d).
[17] Kenneth W. Simons, *Should the Model Penal Code's Mens Rea Provisions Be Amended?*, 1 OHIO ST. J. CRIM. L. 179, 195-96 (2003).

to bring about a prohibited state of affairs, while mere knowledge that it will result from one's conduct is not sufficient. As seen above, treason is one such crime.[18] But there are many others. The crime of falsely incriminating another is defined as "giv[ing] false information to any law enforcement officer with *purpose* to implicate another."[19] It is an offense to harbor or conceal a person if done "with *purpose* to hinder the apprehension, prosecution, conviction or punishment of another for crime."[20] Likewise, it is an offense to "threaten[] unlawful harm to any person with *purpose* to influence his decision, opinion, recommendation, vote or other exercise of discretion as a public servant, party official or voter."[21] Attempt liability requires "*purposely* engage[ing] in conduct that would constitute the crime if the attendant circumstances were as [the defendant] believes them to be."[22] Conspiracy and solicitation require acting with "the *purpose* of promoting or facilitating" the underlying crime.[23]

A natural way to explain why being guilty of these crimes requires the mental state of purpose would be to claim that purposeful misconduct, all else equal, is more culpable than the analogous knowing misconduct. It is axiomatic that punishments must be deserved, and the *amount* of punishment deserved for a crime depends on the seriousness of the crime. Moreover, as Doug Husak points out, "[t]he seriousness of the crime…is partly a function of the culpability of the offender."[24] Thus, if a culpability version of DDE were shown to be defensible, it would explain why it can be sensible to single out some types of purposeful misconduct as meriting harsher punishment than the analogous merely knowing misconduct.[25]

---

[18] *Haupt*, 330 U.S. at 641.
[19] Model Penal Code § 241.5. (emphasis added).
[20] *Id*. § 242.3 (emphasis added).
[21] *Id*. § 240.2 (emphasis added).
[22] *Id*. § 5.01 (emphasis added). *See also* Nelkin and Rickless, *supra* note 10 at 2-4.
[23] *Id*. §§ 5.02, 5.03.
[24] DOUGLAS HUSAK, OVERCRIMINALIZATION: THE LIMITS OF THE CRIMINAL LAW 182 (2008).
[25] Of course, other explanations might be found. Perhaps purposeful misconduct is somehow more serious than the analogous merely knowing misconduct, although the former is no more culpable than the latter. I have doubts about this possibility. Still, my only point is that if we secure a culpability version of DDE, this would *suffice* to make

A third reason to be interested in culpability versions of DDE will emerge in section 1.3, where I argue that certain problems for agent-centered rationales for DDE can be avoided by focusing on culpability, rather than other normative concepts.

1.2 A non-absolutist version of DDE

It is now standard to formulate DDE as a non-absolutist claim.[26] That is, rather than taking DDE to assert that intentionally causing a bad state of affairs always is all-things-considered worse (in the relevant sense) than doing so merely knowingly, DDE more plausibly asserts only that there is *some respect* in which the former is worse than the latter—*e.g*. that it is harder to justify[27] or that there is more *pro tanto* reason against it.[28] I follow this trend in this paper.

To illustrate the difficulty with the absolutist version of DDE, consider the challenge raised by some legal scholars against the distinction between the culpability hierarchy's top two levels: purpose and knowledge. Larry Alexander has argued that purposeful criminal conduct is not "always more culpable than [conduct performed with] knowledge or recklessness."[29] He raises two cases that illustrate the point. For one, Alexander suggests that sometimes "a purposeful criminal actor who imposes a very tiny risk [of harm] is less culpable than a nonpurposeful actor who imposes huge risks for weak reasons."[30] The second example concerns a group of trapped spelunkers who believe they must kill and eat one member of the group to save the rest. Perhaps the spelunkers have a lesser evils justification, which might make their intentional killing seem less culpable than garden-variety knowing killings that are unjustified. But in response, the

---

sense of the practice of punishing purposeful misconduct more harshly than analogous knowing misconduct.

[26] *See* Nelkin & Rickless, *supra* note 10 at 11. Wedgewood also emphasizes the need to focus not on overall justifiability, but rather just on the bad-making features of one's conduct. *See supra* note at 14 at 385-86 (2011). He, in turn, attributes the insight to Quinn. *See supra* note 11.

[27] Nelkin & Rickless, *supra* note 10 at 13.

[28] Wedgwood, *supra* note 14 at 384 (taking DDE to be the claim that there is "a stronger reason against an act if that act has a bad state of affairs…as one of its intended effects than if [it] is merely one of the act's unintended effects").

[29] Larry Alexander, *Insufficient Concern: A Unified Conception of Criminal Culpability*, 88 CAL. L.REV. 931, 943 (2000).

[30] *Id*.

defender of DDE could just restrict her claim that purposeful misconduct is worse than knowing misconduct to cases of *unjustified* action. Nonetheless, Alexander notes, even if the trapped spelunkers *do not* have any justification for their conduct (suppose they in fact needn't kill anyone to survive), they still seem less culpable than "those who impose high risks of death on others *for the mere thrill of it* but who do not have others' deaths as their conscious object."[31] This, then, would be an example of an unjustified, purposeful killing that is on-balance *less culpable* than an unjustified, merely foreseen killing.

Alexander seems correct on this score. But non-absolutist versions of DDE are consistent with Alexander's point that not all cases of purposefully causing harm (or more generally, a bad or wrongful state of affairs) are more culpable than doing so merely knowingly. To formulate DDE so it is consistent with Alexander's point, we should focus not on the *overall* culpability of an action, but rather on the various *contributors* to culpability—*i.e*. the action's culpability-enhancing features. This yields the following non-absolutist, culpability version of DDE (which I'll restrict in one further way shortly):

> DDE_{NAC}: If P1 does A1 intending to bring about a state of affairs that is bad or wrongful to degree X and P2 does an otherwise identical action, A2, not intending, but merely knowing (foreseeing) that it will bring about a state of affairs that is bad or wrongful to degree X, then this is *one respect* in which P1 is more culpable for doing A1 than P2 is for doing A2 (*i.e*. it makes A1 *pro tanto* more culpable than A2).

In other words, A1 has a culpability-enhancing feature, F1, in virtue of being done with the intention to bring about a bad or wrongful state of affairs, while A2 has a *different* culpability-enhancing feature, F2, in virtue of being done while foreseeing that it will bring about such a state of affairs, and F1 adds to the culpability of A1 more than F2 adds to the culpability of A2. This is consistent with A2 still being on-balance more culpable than A1 (provided all else is no longer equal). After all, the *pro tanto* greater culpability that A1 has in virtue of possessing F1

[31] *Id.* (emphasis added).

rather than F2 can be outweighed by other culpability-enhancing features of A2, which A1 does not possess. For example, perhaps A1 is on-balance justified while A2 is not. Or perhaps P1 is aware of some additional considerations that justify A1 *somewhat* but not fully, while these considerations do not apply to A2 at all. This would be a case where both A1 and A2 are unjustified, but A2 is still on-balance more culpable than A1.[32] DDE_{NAC} is consistent with such cases, and so it accommodates Alexander's point that sometimes purposefully harming is on-balance less culpable than doing so only knowingly.[33]

1.3 Why agent-centered rationales are of particular interest for the criminal law

Now that we have a plausible culpability version of DDE in view, we can ask which rationale for it seems most promising—an agent-centered or a victim-centered rationale. Victim-centered rationales are becoming increasingly popular.[34] For instance, Nelkin and Rickless argue that "the best rationale for [DDE] is the 'means' rationale, according to which it is wrong to use others as means without their consent."[35] The means rationale is victim-centered in that it is "grounded in the fact that persons have a right not to be caught up, to their disadvantage, in the harmful direct

[32] To illustrate, notice that the following pair of cases is no counterexample to DDE_{NAC}. In Case 1, P1 intentionally kills innocent victim, V, in order to save P1's son from being tortured for an afternoon. An evil tyrant will refrain from torturing P1's son iff V dies by P1's hand. Thus, P1 shoots V intending to cause V's death. Suppose that this killing isn't actually justified, although it nearly is. In Case 2, a trolley is careening down the track towards a four-leaf clover, and to prevent it from getting squished, P2 redirects the trolley onto another track to which V is bound. P2 acts while merely knowing that he will cause V's death. Still , P2's act is *much* more culpable than P1's act.

DDE_{NAC} isn't undermined by these cases. This is because P1 perceives additional reasons that help justify his action (although they don't fully justify it), but these considerations don't apply to P2. While there is one respect in which P1's action is *more culpable* than P2's (in virtue of P1 intending the death and P2 only foreseeing it), this is outweighed by another difference between them—namely, that the benefits of P1's action are so much greater than P2's action. Thus, P2's action is *on-balance far worse* than P1's. DDE_{NAC} is compatible with this result because it just entails that there is *one respect* in which P1's act is worse than P2's.

[33] One might object that because DDE_{NAC} is not formulated in terms of on-balance culpability, it is too weak to be interesting. Nonetheless, it has implications about overall culpability *provided all else is equal*. Specifically, if we hold fixed the both the harm in question and the benefits the two actors are seeking, then we plausibly can say that intending harm is on-balance worse than merely foreseeing it. DDE_{NAC} can be formulated accordingly.

[34] Nelkin & Rickless, *supra* note 10 at 11-13; Nelkin & Rickless, *So Close, Yet So Far: Why Solutions to the Closeness Problem for the Doctrine of Double Effect Fall Short*, 49 NOUS 376, 404 (2015); Nelkin & Rickless, *Three Cheers for Double Effect*, 89 PHILOSOPHY AND PHENOMENOLOGICAL RESEARCH 125, 128 (2014); Quinn, *supra* note 11; Wedgewood, *supra* note 14.

[35] Nelkin & Rickless, *supra* note 10 at 17; *see also So Close, Yet So Far*, *supra* note 34 at 402-04.

agency of others."[36] Although I will defend a version of the agent-centered rationale, I don't mean to suggest that the means rationale isn't attractive. Quite the contrary. In fact, I think my agent-centered rationale remains available as a supplement to the means rationale. It seems possible to be a pluralist about the grounds for DDE.

Nonetheless, there are reasons to be dissatisfied with the means rationale if one's interest is explaining and justifying existing criminal law. The means rationale *by itself* appears unable to fully justify the way the criminal law actually encodes DDE. After all, the law often takes intentional wrongdoing to constitute a more serious crime, meriting harsher punishment, than the analogous merely knowing wrongdoing—and this is so *even* in cases like *Haupt* where no one is obviously used as a means without their consent. One might object that the father in *Haupt* really did use his son as a means to his objective of aiding an enemy of the United States. But even if this is the case, there is no indication that his son withheld his *consent* to be involved in this way in his father's plan (which is central to Nelkin and Rickless's explanation of the means rationale). After all, the son appears to have willingly accepted his father's help getting housing, a job and a car. What's more, treason can be committed in ways that do not involve using anyone as a means to aid the enemy—*e.g*. by stealing computer passwords or technical schematics and then passing them along to the enemy in order to harm the U.S.[37] This problem is, of course, not limited to the crime of treason.[38]

This obstacle will come as no surprise to proponents of the means rationale like Nelkin and

[36] Nelkin & Rickless, *supra* note 10 at 17.
[37] Other accounts of the means principle might treat using the property of others as analogous to the use of the person. I doubt that this analogy can be sustained, but I can't deal with all forms of the means principle here.
[38] Other purpose crimes can be committed in ways that do not use anyone as a means without their consent, or do not otherwise directly involve anyone else. For example, concealing a fugitive is criminal if done "with purpose to hinder [his or her] apprehension…" *See supra* note 20. But it is possible to commit this crime even if everyone involved in the act of concealment consents to being used that way. Moreover, one might conceal someone for this purpose without directly involving the concealed person—*e.g*. by planting misleading information about the fugitive's whereabouts in the police computer system. Similar points apply to conspiracy and solicitation. *See supra* notes 22-23.

Rickless. They explicitly acknowledge that the means rationale "does not fit perfectly with the general distinction between intended and merely foreseen harm,"[39] which after all is the distinction employed in the criminal law. Nonetheless, it seems the means rationale must be supplemented if we are to fully justify the law's use of DDE.

Accordingly, those seeking to explain and justify existing criminal law have reason to take a closer look at agent-centered rationales. But the agent-centered rationale faces problems of its own. For one, the idea that it is especially bad to "aim at evil" is imprecise (though my argument will help mitigate this problem). More importantly, Nelkin and Rickless argue that "defenders of the aiming-at-evil rationale are caught on the horns of a dilemma depending on how they choose to understand the nature of evil."[40] First, "evil" might be understood not merely as bad states of affairs, but rather as wrongful ones. However, "if wrongness is part of the essence of evil, then it is circular to explain the wrongness of an action (or its tendency to be wrong) by adverting to the fact that, in performing the action, the relevant agent aims at something that is wrong."[41] On the other hand, if "evil" is understood as states of affairs that are bad, we encounter the distinct problem that "it does not seem wrong in itself to aim at something very bad (such as great harm): for example, it does not seem wrong in itself to aim at harming people who were or are engaged in wrongful attacks on other people."[42]

However, defenders of the aiming-at-evil rationale can in fact avoid both horns of this dilemma by employing certain resources suggested by Michael Moore's discussion of DDE in the criminal law.[43] The first horn can be avoided by formulating DDE not in terms of wrongness, but rather in terms of *culpability*—as Moore does and I have already endorsed doing. This

[39] Nelkin & Rickless, *So Close, Yet So Far*, *supra* note 34 at 402.
[40] *Id*.
[41] *Id*.
[42] *Id*.
[43] *See* Moore, *supra* note 15 at 48. Moore focuses on a culpability version of DDE that concerns only cases in which there is no justification for bringing about the bad or wrongful state of affairs in question.

sidesteps the first horn because it's not viciously circular to explain why intentionally bringing about a state of affairs, S, is more *culpable* than doing so only knowingly by appeal to the fact that S is wrongful and should be avoided. This is yet another reason to formulate DDE in terms of culpability. What's more, the second horn of the dilemma can be avoided by following Moore in focusing on cases where the relevant bad or wrongful state of affairs, S, one brings about is not otherwise *justified—e.g.* is not outweighed by any countervailing good one seeks to bring about by way of S. After all, aiming at a state of affairs that not only is bad or wrongful, but that also is unjustified, does seem wrong in itself—*i.e*. an independent source of culpability.

My agent-centered argument for a culpability version of DDE follows Moore's suggestions. It concerns culpability and primarily applies to cases of *unjustified* harming. One might want to extend the argument to cases in which the intended or foreseen harm is justified—as in classic trolley cases, where one person is killed to save five. But I am not confident that the implications of DDE_{NAC} are plausible when it comes to cases involving justified harming (since I am not sure one can be culpable for acting in ways that one reasonably sees as justified).[44] Accordingly, I focus on cases where the relevant harming is unjustified, as can be expected in criminal cases (where no affirmative defenses apply). If my argument succeeds, it will only establish a restricted version of DDE_{NAC}, *viz*. one that is limited to scenarios where the relevant harm is unjustified. That, I think, is sufficient for the project of finding a normative foundation for the criminal law's use of DDE. Thus, the restricted version of DDE I will defend is this:

> DDE_{NACR}: If P1 does A1 intending to bring about a state of affairs that is bad or wrongful to degree X, and P2 does an otherwise identical action, A2, not intending, but merely knowing (foreseeing) that it will bring about a state of affairs that is also bad or wrongful

[44] *If* one wants to say that it's more culpable to intentionally kill one in order to save five than it is to merely foresee that one will die as the result of saving five, then we could attempt to extend the argument I offer below to explain this result. However, I'm not sure this is what we want to say about the cases. Since both cases plausibly involve conduct that the actors know to be justified and thus permissible, one might think neither actor is *at all culpable* for his conduct. Accordingly, I suspect there is independent reason to follow Moore in restricting our culpability version of DDE to cover *only cases involving unjustified conduct*.

to degree X—*and there are no independent considerations sufficient to justify bringing about these states of affairs*—then this is one respect in which P1 is more culpable for doing A1 than P2 is for doing A2 (*i.e*. it makes A1 *pro tanto* more culpable than A2).

One might worry that DDE_{NACR} loses something that is crucial to proponents of DDE. Since DDE_{NACR} applies only when the act in question is unjustified, DDE_{NACR} plays little or no role in determining whether acts are justified, and thus permissible. But being able to play such a role, one might think, is an important feature of DDE. Be that as it may, my aim is only to defend a version of DDE that can do what is needed for purposes of justifying the criminal law. Even if DDE_{NACR} only applies when the actor has no available justification (*i.e*. no affirmative defense), it still would suffice to vindicate standard criminal law practice. After all, it would help explain why defendants who intend a prohibited result should be subject to harsher penalties. Thus, while DDE_{NACR} perhaps can't give all proponents of DDE everything they want, it would still give criminal law theorists everything they need.

1.4 A lingering challenge: the closeness problem

The agent-centered argument for DDE_{NACR} I offer below will not solve *all* the challenges that such approaches to DDE face. In particular, the so-called *closeness problem* remains. It arises because of the difficulty in distinguishing effects that are intended from effects that are merely foreseen.[45] DDE_{NACR} presupposes that some such distinction is tenable. However, many cases where the actor seems to intend a bad state of affairs might plausibly be redescribed so that this state of affairs now seems like a merely foreseen side-effect. If Jerry pushes George onto the tracks to prevent a trolley from killing five innocent people further down the tracks, it might initially seem plausible that Jerry intends to harm George. Nonetheless, Jerry might reply that he does not in fact intend to harm George—he merely intends to push George onto the tracks and thereby stop the trolley. Jerry would greatly prefer it if this neither killed nor harmed George.

[45] *See* Wedgwood, *supra* note 14 at 393-94.

Since many typical cases of intended harm admit of this sort of redescription, the category of intended bad effects seems so "close" to the category of merely foreseen effects that the distinction between the two threatens to collapse.

I will not attempt a complete solution to this problem here. Some solutions have been suggested,[46] though others are skeptical that a solution can be found.[47] Ultimately, the defensibility of DDE_{NACR} depends on solving the closeness problem. Still, let me try to mitigate the force of the problem in three ways.

First, *contra* Nelkin and Rickless, the closeness problem seems no more problematic for the agent-centered rationale, and associated versions of DDE that use the intend/foresee distinction, than the analogous difficulty is for the means rationale. Nelkin and Rickless argue that a version of DDE that conforms to the means rationale avoids the closeness problem, and therefore is preferable to versions of DDE that track the intend/foresee distinction.[48] Following Quinn, Nelkin and Rickless distinguish between *harmful direct agency*, "in which harm comes to some victims, at least in part, from the agent's deliberately involving them in something in order to further his purpose precisely by way of their being so involved," and *harmful indirect agency*, "in which harm comes to some victims, but in which either nothing in that way is intended for the victims or what is so intended does not contribute to their harm."[49] Nelkin and Rickless then endorse a version of DDE on which harmful direct agency is harder to justify than equally harmful indirect agency.[50] This principle avoids the closeness problem, they claim, "because it does not require that harm itself be intended for the behavior to be in the disfavored category."[51]

[46] *See* FitzPatrick, *supra* note 13 at 187; Wedgwood, *supra* note 14 at 396-97; M. Liao, *The Closeness Problem and the Doctrine of Double Effect: A Way Forward*," CRIM. L. & PHILOSOPHY (forthcoming).
[47] Nelkin & Rickless, *So Close, Yet So Far*, *supra* note 26.
[48] Nelkin and Rickless, *supra* note 10 at 11-13.
[49] *Id*. at 13; *see also So Close Yet So Far*, *supra* note 26 at 404.
[50] Nelkin and Rickless, *supra* note 10 at 13.
[51] *Id*.

Nonetheless, their proposal faces an analogous challenge: namely, distinguishing harmful direct agency from harmful indirect agency. This distinction depends on the intentions of the agent (which is crucial to securing their argument that intentions matter to wrongdoing). However, to draw this distinction, we need to be able to decide whether the agent intends to directly involve others in his plans—*i.e*. whether the agent intends to use others for his own purposes—or whether "nothing is in that way intended for the victims"—*i.e*. whether the agent does not intend, but merely foresees that others will be affected. Thus, while Nelkin and Rickless are correct that their version of DDE "does not require that harm itself be intended for the behavior to be in the disfavored category," it still requires that *the use of others* be intended, rather than merely foreseen, in order to count as harmful direct agency. Accordingly, Nelkin and Rickless's version of DDE, premised on the means rationale, seems to require meeting a challenge very close to the closeness problem. As a result, DDE_{NACR} and my agent-centered rationale are no worse off on this score than versions of DDE grounded in the means rationale.

Hopefully, I can also do more to alleviate the force of the closeness problem for present purposes. Specifically, the second step I want to take in this direction is to focus on pairs of cases in which the sort of redescription that fuels the closeness problem is not plausible. Consider this pair of cases where the problematic redescription seems unavailable:

> ***Arson 1***: Tony, a mob boss, offers to pay Alan $5000 to burn down a building, but it has to be done before midnight or Alan won't get paid. Alan agrees. He arrives at the building at 11:30pm, and as he is about to light the fire, he sees that Victor is doing something on the second floor. Victor does not leave, so Alan proceeds to light the fire knowing (*i.e*. while practically certain) that this will lead to Victor's death. As expected, Victor dies. Alan finds it regrettable that Victor dies, but he decides there's nothing he could do—he "really needed the money."
>
> ***Arson 2***: This case is as similar to Arson 1 as can be, except that now Tony offers to pay Bobby $5000 to see to it that Victor dies tonight before midnight. Moreover, Bobby is to kill Victor by making it look like he was killed by a fire in the building. Tony will not give Bobby the money unless Victor actually dies. (Suppose Tony has a perfectly reliable method for determining this.) Just before midnight, Bobby lights the building on fire and Victor dies.

Bobby finds it regrettable that Victor dies, but he decides there's nothing he could do—he "really needed the money."

Arson 1 is a merely foreseen (knowing) killing, while Arson 2 is a case in which the killing is intended—specifically, intended as the means to obtaining the $5000. Bobby acts with the purpose to bring about Victor's death, while Alan merely knows that his act will cause the death. Moreover, in Arson 2, since Victor's death is *necessary* for Bobby to get paid, the case cannot plausibly be re-described as one in which Bobby does not intend Victor's death, but merely foresees it. Thus, Arson 1 & 2 is a pair of cases where the distinction between intended harm and merely foreseen harm appears to be relatively stable, and not liable to collapse.[52]

Granted, the difference between these cases does not matter from the legal perspective, since both Alan and Bobby would be guilty of murder, a crime that requires only knowledge. Nonetheless, to keep things simple as possible, I use Arson 1 & 2 as my main example in what follows. It is easy enough to construct similar pairs of cases using various purpose crimes. For example, we get an analogous pair involving treason—where the purpose/foresight distinction *does* have legal (not just moral) significance—simply by replacing the state of affairs foreseen in Arson 1 and intended in Arson 2 (*i.e*. death) with a state of affairs in which the enemy is aided.[53]

Arson 1 & 2, then, represent the best sort of cases to focus on. In many real-life cases, the distinction between intended and merely foreseen effects will not be this stable. Often, the bad

[52] One might worry that the closeness problem still arises even in this pair of cases. After all, in Arson 2, Bobby might insist that he didn't actually intend to kill Victor, but rather only intended to make it *appear* as though Victor were dead, so he would receive the $5000. But this worry is blocked by certain stipulations about the case. Specifically, we're supposing that Tony has a *perfectly reliable* method of detecting whether Victor in fact is dead. Thus, Tony cannot be fooled into thinking that Victor dies when in fact he survives. Given this stipulation, Bobby will have to aim at Victor's death itself. Bobby will be unable to maintain that in fact he only aimed to make it appear as though Victor died.

[53] Here is the treason analog of Arson 1. Suppose that Fritz, an operative for the Axis powers during WWII, offers to pay Alan $5000 to burn down a military building by starting an electrical fire, but it has to be done before midnight or Alan won't get paid. Alan agrees. He arrives at the building at 11:30pm, and as he is about to light the fire, he sees that the building contains a shipment of technical schematics whose destruction would hurt the US war effort. No one arrives to cart away the technical schematics before midnight, so Alan proceeds to light the fire knowing that this will end up aiding an enemy of the U.S. As expected, the schematics are destroyed in the fire. Alan finds it regrettable that he thus aided the enemy, but he decides his hands were tied—he "really needed the money." (It is now easy to see how to formulate a comparable version of Arson 2 that involves treason.)

state of affairs one brings about will not *itself* be necessary for obtaining the actor's goal. Rather, it will often be the case that what's useful as a means to one's goal just is something closely *connected* to the bad state of affairs in question (*e.g*. burning down a building, redirecting a trolley), not the bad effect itself (*e.g*. death).

Nonetheless, any argument for some version of DDE will have to make sense of these simpler cases *as well*. Thus, the argument I offer below in favor of DDE_{NACR} will at least account for straightforward cases like Arson 1 & 2 (and their treason analogs, *etc*.). Moreover, supposing we eventually figure out how to draw a clear-cut distinction, applicable even to real-life cases, between genuinely intended and merely foreseen effects, the hope is that my argument will still go through. In this way, I aim to offer a rationale for DDE_{NACR} while sidestepping (or postponing) the need to provide a general solution to the closeness problem.

The third mitigating point I want to make is that the substance of the argument offered below for DDE_{NACR} will end up giving some practical guidance for how to proceed in real life cases where the distinction between intention and foresight is unclear. Specifically, it indicates what sort of *evidence* to look for to decide whether a given bad effect is intended or merely foreseen. I return to this point in Section 5.

2. An Agent-Centered Rationale for the Criminal Law's Version of DDE

DDE_{NACR} entails that there is one respect in which Bobby is more culpable for his conduct in Arson 2 than Alan is for his conduct in Arson 1—perhaps not much, but a bit. Because there are no other morally relevant considerations, this means Bobby's conduct is on-balance slightly more culpable than Alan's. Some might not be convinced that these implications are plausible. The difference in culpability between the two cases, if any, is only slight. Nonetheless, I want explain why DDE_{NACR} might seem plausible even as applied to cases like Arson 1 & 2. This,

after all, would help justify the criminal law's reliance on something like DDE_{NACR}. Accordingly, I am engaged in an attempt at rational reconstruction: something like the argument below would make sense of the law's use of DDE.

I begin by sketching my defense of DDE_{NACR}, which I refine in section 3. The argument is based on the thought that the main difference between i) intending bad or wrongful states of affairs—which for simplicity I henceforth refer to as intended *harms*—and ii) merely foreseen ones is a matter of one's *commitment* to them. For intended harms, one is at least somewhat committed to bringing them about under the circumstances, while merely foreseen harms involve no such commitment.[54] Thus, the key difference between Arson 1 & 2 is that Alan is not committed to causing Victor's death at all, while Bobby is. That is, Bobby, but not Alan, will feel at least some motivational pressure to take steps that seem to promote the death.

Alan's lack of commitment can be cashed out in terms of the motivational pressures he would experience in various choice scenarios—*i.e.* what I call his *motivational profile*. One way Alan's lack of commitment to Victor's death would come to light concerns the extent to which he'd be motivated to perform variations of his actual conduct, which differ in terms of whether they make Victor's death more likely. If Alan knows that lighting the fire in the west wing of the building would make Victor's death more likely than lighting it in the east wing, then because Alan is not committed to the death, he would not feel any inclination to select the first option over the second (assuming there are no other relevant considerations). By contrast, because Bobby *is* committed to Victor's death, he *would* feel at least some motivational pressure to light the fire in the west wing of the building in this variation of the case.

[54] Others understand the difference between intending an effect and merely foreseeing it in precisely this way. *See* MICHAEL BRATMAN, INTENTION, PLANS AND PRACTICAL REASON 141-42 (1999) (arguing that intending an effect entails being committed to it in three ways, while merely foreseeing the effect does not); Allison Hills, *Defending Double Effect*, 116 PHILOSOPHICAL STUDIES 133, 134 (2003) (observing that "an agent intends some state of affairs if she is committed to bringing it about," but "merely foresees it" if there is no such commitment, where "commitment" is understood in essentially the same way I construe it here).

A second way in which Alan's lack of commitment makes his motivational profile different from Bobby's has to do with the *further* steps they would be inclined to take to ensure that Victor dies. Suppose that Victor miraculously does *not* die in the fire. Given that Alan is not committed to the death, he would not be inclined to take additional steps to see to it that Victor really dies. By contrast, in Arson 2, if Victor does not initially die from the fire, Bobby *would* feel some inclination to take further steps to ensure that he dies (at least assuming Bobby does not change his mind, become irrational, *etc*.). After all, in Arson 2, for Bobby to achieve his aim of getting the $5000 for killing Victor, it really is necessary that Victor dies.

More generally, I claim that when one intends a harm, either as an end or as a means, one is actually somewhat committed to it in this sense:

> *Commitment*: P is somewhat committed to bringing about a state of affairs, S, iff P's motivational profile is such that, in scenarios where P's ends remain the same as they actually are and S retains the same instrumental or constitutive relationship to those ends that S has in the actual world, P would experience *at least some motivational pressure* to seek out and take steps that appear to him to *promote*[55] the occurrence of S (*e.g*. by making S more likely than if those steps had not been taken)—at least provided those steps are not too costly, P does not suffer from execution-failure or irrationality, and so on. The broader the range of cases in which P would feel motivational pressure to take steps that promote S, and the stronger this felt motivational pressure is, the *more committed P is* to bringing about S.

While I suggest that intending harm makes one at least somewhat committed to bringing about the harm in this sense, the same is not true for merely foreseen harm.

One clarification. Both Alan and Bobby are similar in this respect: if there were an *alternative* course of conduct they could take that would easily secure the desired $5000, but that would *not* result in Victor's death, they both would take that action instead. This is because Alan is not committed to the death at all, and Bobby does not intend the death as an end in itself, but

[55] I understand the promotion relation as Schroeder describes it: "X's doing A promotes p just in case it increases the likelihood of p relative to some baseline. And the baseline, I suggest, is fixed by the status quo." MARK SCHROEDER, SLAVES OF THE PASSIONS 113 (2007). Here, the relevant baseline is S's likelihood of occurring in a case that is exactly like the actual world in which the contemplated additional steps are not taken—*i.e*. in which one's action A is performed but the extra steps are not.

rather only as the means to getting $5000. Thus, not even Bobby is committed to the harm in all possible worlds. Nonetheless, this is consistent with Bobby's being committed to the harm in the above sense. After all, a case in which Bobby can obtain the $5000 without killing Victor—call it *Free Money*—is *not* one in which the killing is instrumentally related to Bobby's ends in just the same way as it is in the actual scenario, *i.e.* Arson 2. Accordingly, my claim here does *not* entail that Bobby would feel any motivational pressure to take steps to ensure Victor's death in *Free Money*. Rather, my only claim is that in cases where the killing *does* remain instrumentally necessary to Bobby's getting paid the $5000, just as it is in the actual world (*i.e.* Arson 2), Bobby *would* feel motivational pressure to take steps to make Victor's death more likely. That is the sense, I claim, in which Bobby is committed to Victor's death, but Alan is not.

Accordingly, what seems to underlie DDE_{NACR} is the idea that there is something especially culpable about being committed to causing a given harm in the way that Bobby is in Arson 2. But it will not do to merely *assert* that such a commitment renders one more culpable. Some explanation is needed.

The explanation I propose employs a familiar theory of culpability. Its basic thought is that one is culpable for an action to the extent it *manifests insufficient regard* for the interests of others (or perhaps more generally, for morally relevant interests).[56] More precisely:

> *Insufficient Regard Theory*: P is culpable for her action, A, to the extent that A manifests P's insufficient regard for morally relevant interests—or equivalently, the extent to which A demonstrates that P has failed, in her motivational processes, to give due weight to the normative reasons she actually has (and has access to based on her evidence) in the circumstances.

[56] *See, e.g.*, Larry Alexander, *Culpability*, in THE OXFORD HANDBOOK OF PHILOSOPHY OF CRIMINAL LAW (Deigh, and Dolinko, eds.) (2011) ("acts are culpable when they manifest insufficient concern for the interests of others"); LARRY ALEXANDER & KIMBERLY FERZAN, CRIME AND CULPABILITY 67-68 (2009) (arguing that "insufficient concern [is] the essence of culpability"); NOMY ARPALY & TIM SCHROEDER, IN PRAISE OF DESIRE 170 (one is blameworthy for A "to the extent that A manifests ill will (or moral indifference) through being rationalized by it"); Peter Westen, *An Attitudinal Theory of Excuse*, 25 LAW & PHILOSOPHY 289, 373-74 (a person is "blameworthy for…conduct that a statute prohibits if he was motivated by an attitude of disrespect for the interests that the statute seeks to protect," e.g. "malice, contempt, indifference, callousness, or inadvertence").

Thus, if you do some action that on-balance there is reason not to do, and you reasonably could and should have appreciated that the reasons against the action outweigh the reasons in favor of it, then you have acted in a way that manifests your insufficient regard for morally relevant interests (unless excused). Moreover, this view incorporates a certain *principle of lenity*, which will become crucial when I refine the argument later. I will explain the point in more detail then. But for now, the thought behind this principle of lenity is that the amount of insufficient regard *manifested* by an action is equal to only the *minimum* amount that it is necessary to postulate in order to explain why the actor behaved as she did under the circumstances. (More on this below.[57]) Although my argument is formulated in terms of the insufficient regard theory in order to fix concepts and enhance clarity, this should not make the argument overly controversial. For this already prevalent theory is similar in spirit to a range of other theories of culpability.[58] Thus, I suspect my argument can be translated into the terminology of other theories (although I will not argue for that here). Strictly speaking, though, my argument depends on some version of the insufficient regard theory, or a related theory, proving true.

Now, the rationale for DDE_{NACR} is that intending, and thus being committed to causing, a given harm manifests more insufficient regard for others, all else equal, than merely foreseeing that one's action will cause that harm, without being committed to it. Why might this be so?

The basic idea is that when you know, or foresee, that a particular action of yours, A, will cause a given harm, then your conduct manifests insufficient regard in one way, while when you intentionally cause harm, and thus are committed to it in the above sense, your conduct manifests

[57] *See infra* note 74 and accompanying text.

[58] The insufficient regard theory is similar in spirit to the theory 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." GIDEON YAFFE, ATTEMPTS 38 (2011). *See also* Julia Markowitz, *Acting for the Right Reasons*, 119 PHILOSOPHICAL REVIEW 201 (2010) (discussing the theory that "my action is morally worthy if and only if my motivating reasons for acting coincide with the reasons morally justifying the action"); Peter Graham, *A Sketch of a Theory of Blameworthiness*, 88 PHILOSOPHY AND PHENOMENOLOGICAL AFFAIRS 388, 407 (2014) (arguing that one is blameworthy iff the reactive emotions are appropriate, and this, in turn, is true iff "in ϕ-ing, X has violated a moral requirement of respect").

insufficient regard in this way *as well as a second way*. In particular, in the case of merely foreseen harm, you fail to be sufficiently repelled by the badness of the harm. That is, the harm you know your action A will cause is a normative reason against A, but you do not have a sufficiently strong motivating reason against A in virtue of knowing that A will cause the harm. Thus, your motivating reasons do not coincide with your actual reasons: you failed to have a motivating reason that you should have had, or at least one of the required strength. By contrast, when you do A with the intention to bring about the harm in question, and thus are committed to it in the above sense, you manifest insufficient regard in the same way as merely foreseen harm, *as well as* in an additional way. Beyond being insufficiently repelled by the harm, you also display the further fault of taking it that there is a positive reason in favor of promoting the harm. That is, your act demonstrates that promoting the harm is something to which you are affirmatively attracted more than you ought to be, assuming the harm is unjustified. Since acting with the commitment to harm involves *two* manifestations of insufficient regard, while merely knowing or foreseen harm involves only *one*, the insufficient regard theory entails that there is a respect in which the first actor will be more culpable for his conduct than the second is for hers.

3. A More Precise Defense of the Argument

This, then, is the basic line of thinking that I claim supports DDE_{NACR}. However, the argument must be refined in light of a number of complications. I begin by clarifying the picture of motivation I am working with. Plausibly, when you do an action, A, you take it that there are certain reasons in favor of A—call them "R+"—and you may or may not take it that there are certain reasons against A—call them "R-." I am supposing that to "take it" that there is a reason in favor of A is to experience some *motivational pressure* towards doing A, while to "take it" that there is a reason against A is to feel somewhat motivationally repelled from doing it. Thus,

when you do A, this shows (assuming you are rational) that your motivational pressures towards A outweigh the felt motivational pressures against A. In other words, it shows that you regard it as true that R+ > R-. But if in fact R- > R+ (and you reasonably can and should know this), then your doing A anyway manifests insufficient regard for morally relevant interests. That is, it demonstrates that you do not attach sufficient weight to the reasons against the action (*i.e.* sufficient compared to the weight you reasonably should attach to the reasons in favor of the action[59]). Likewise, in talking about "attaching" weight a consideration, I do not mean merely your *beliefs* about how much it counts for or against the action. Rather, it is a matter of experiencing some motivational pressure towards or against the act in virtue of believing the relevant consideration.[60] (I take it to be an open question whether you experience this motivational pressure in virtue of your desires or in virtue of your evaluative beliefs. That is, my argument is compatible with, but is not supposed to require, a Humean theory of motivation.)

Given this picture, let me precisely state my argument—I call it the "Two Strikes Argument"—before defending each of its premises.

The Two Strikes Argument:

1) If you do A with the intention that A help bring about a harm (or more generally a bad or wrongful state of affairs)—call it H for short—then you act with at least some *commitment* to H (in the sense defined above).
2) If you do A while merely foreseeing (knowing) that it will cause an identical harm H, then in doing A you tolerate, but are not committed to H.

[59] In general, I don't think that what matters to culpability is the absolute magnitude of the weight you attach to the reasons against your action. That is, it doesn't matter how strongly, in absolute terms, you feel motivational pressure against the action in virtue of the facts counting against it. Rather, what matters is how much weight you attach to the reasons against the action *as compared with* the weight attached to the reasons in favor of it. More on this below.

[60] Note that my claim here is only that one "regards it as true" that R+>R- iff one does A *provided one is rational and does not suffer from execution failure or the like*. Still, this does not render my notion of "regarding it as true that R+>R-" is empty or explanatorily useless. On my view, "regarding it as true that R+>R-" means "feeling more motivational pressure in favor of A than against it." But this does *not* mean the same as "doing A (provided one is rational)." First, these terms clearly have different intensions, since the first concerns one's motivational processes, while the latter concerns behavior. Moreover, the two terms are not extensionally equivalent either. I can regard it as true that R+>R- in the sense that I feel more motivational pressure in favor of A than against it, but still fail to do A even though I'm rational. For instance, it's possible that just as I am about to do A, a mad scientist paralyzes my body. Cognitively, I am exactly as I was before being paralyzed—I remain fully conscious and have all the same mental states. Thus, I'm still rational. The paralysis simply blocks the execution of my intention. Thus, "regarding it as true that R+>R-" is also not co-extensive with "doing A (provided one is rational)."

3) If you do A with a *commitment* to H, then you display the following two distinctive faults provided neither your action A nor H itself is justified:
   a) you are insufficiently repelled by H, and
   b) you feel some motivational pressure to affirmatively promote H (*i.e*. adjust your conduct to make H more likely, take further steps to ensuring it, opt for alternatives that make it more likely, *etc*.), and this is more motivational pressure towards promoting H than you *should* experience.
4) By contrast, if in doing A you only tolerate but are not committed to H, then your conduct only manifests *one* of the two faults in 3): *i.e*. that you are insufficiently repelled by H (and insofar as you do happen to feel some attraction to H, this will not be *manifested* in your action in a way that bears on your culpability for A).
5) Thus, if you do A intending that it bring about H, then (assuming A and H are not justified) there is one respect in which A is more culpable than doing A while merely foreseeing that it will cause H (*i.e*. the former has a culpability-enhancing feature the latter lacks).

Premises 1) & 2)

I take premises 1) and 2) to be conceptual truths. For those who are skeptical of defending claims by calling them conceptual truths, 1) and 2) can be thought of as stipulative claims about how I understand the distinction between intending a state of affairs versus merely foreseeing it. 3) and 4) are the meat of the argument anyway.[61]

Still, I don't think 1) and 2) should be controversial.[62] The claim in 1)—that acting with the intention to bring about some state of affairs means one is committed, in the sense sketched in section 2, to that state of affairs (*e.g*. by behaving in ways that one sees as tending to promote it)—is similar in spirit to other accounts of the role intentions play in action. Others argue that intentions constitute commitments of various sorts. For example, Gideon Yaffe takes it that intending to X involves, *inter alia*, a commitment in the sense that one would be "criticizably irrational for failing to form an intention for, and thus a commitment to the occurrence of, those conditions that one believes to be necessary for the occurrence of X." [63] The sort of commitment

[61] Even if the intending/foreseeing distinction does not perfectly line up with the distinction between commitment and its absence, I could still argue that the latter distinction is the one the criminal law *should* use.
[62] Others endorse the view in 1) and 2), namely that intending an effect involves a commitment to it, while merely foreseeing it does not. *See* Hills, *supra* note 54 at 134; Alison Hills, *Intentions, Foreseen Consequences and the Doctrine of Double Effect*, 133 PHILOSOPHICAL STUDIES 257, 260 (2007); Bratman, *supra* note 54 at 140-43 (1999).
[63] Yaffe, *supra* note 58 at 82-3. *See generally id*. at 82-90.

I am concerned with here is an even weaker one than this. I take it that intending harm entails a commitment just in the sense that one will experience *some motivational pressure* to take apparently available steps to promote the harm in question, even if one actually does not decide to take those steps for other reasons (*e.g*. they are too costly). Thus, I think it's plausible that this is *one* type of commitment one has when acting with an intention to harm. This is hardly a complete defense of premise 1), but fully defending it would require too much of a detour into philosophy of action.

Similarly, I take it that 2) embodies a conceptual truth about merely foreseen harm. It seems that precisely what distinguishes actions done while merely foreseeing that they will cause harm from actions done with the intention to bring about such harm is that the former do not involve a commitment to the harm (in the above sense), while the latter do.[64] Instead, acting while merely foreseeing that it will result in harm shows only that one is willing to tolerate the harm in order to obtain the benefits of the action. Hence 2).

Premise 3a)

The argument's more substantive claims are premises 3) and 4). Start with 3a). Supposing you are indeed culpable for doing A—as would be the case when both A and the harm in question are clearly unjustified—this will be at least in part because you fail to be sufficiently repelled by the harm itself. If neither A nor the harm you are committed to causing thereby are reasonably seen as justified, then the harm in question should give you a motivating reason to refrain from doing A (*i.e*. it belongs in R-). But since you perform A anyway, this shows (assuming you're rational) that the harm did not repel you enough to get you to abstain from A. Accordingly, since neither A nor the harm are justified, we know you attach too little weight to the reason against A that exists in virtue of the harm you're aware it will cause. Hence 3a).

[64] Again, Bratman and Hills agree. *See supra* notes 54 and 62.

To this, one might object that perhaps someone who is committed to bringing about a certain unjustified harm can really be *sufficiently* repelled by the harm. Suppose, for instance, you ought to be repelled by the intended harm to a certain degree—say, to degree 5 (to arbitrarily assign a number to it). Moreover, suppose you actually feel 5 units of repulsion to the harm. Nonetheless, you wrongly attach slightly greater weight to the reasons you take there to be in favor of the harm (which stem from the benefits the harm is the means to). Thus, you feel 6 units of attraction towards the harm and therefore are committed to bringing it about. But since you feel the required amount of repulsion to the harm (*i.e*. 5 units), wouldn't it be false to say that you are insufficiently repelled by it? What guarantee is there that someone who is committed to bringing about an unjustified harm will be insufficiently repelled by it—as opposed to merely being overly attracted to the perceived benefits of the harm?[65]

This objection fails because it relies on a mistaken picture of what it means to be insufficiently repelled by a particular consideration. The objection assumes that there is some *absolute level* of repulsion that one ought to feel towards a given harm. This is implausible because some actors are likely to have stormy inner lives and experience the various motivational pressures they are subject to much more strongly than other actors with more tranquil inner lives. Accordingly, it seems doubtful that there is an absolute level of strength or intensity of repulsion that one ought to feel towards particular harms. Instead, the level of repulsion one ought to feel towards a harm will depend in part on the strength of the other motivational pressures one is subject to. Suppose you all-things-considered ought not to do A because it would cause grave harm. If you are somewhat attracted to the harm because you think it would also lead to a modest benefit (say 2 units), then *less* motivational repulsion to the harm will be required to get you to behave as you ought than would be required in a different case

[65] Thanks to Jon Quong for pressing me on this objection.

where you think the harm would lead to greater benefits (say, 3 units). Thus, the amount of motivational repulsion required to get you to do what you ought to—*i.e.* that counts as *sufficient*—depends on the other considerations you feel the pull of.

Thus, insufficient repulsion should be understood comparatively. Suppose the harm you are aware your action, A, will cause *should* give you a sufficient motivating reason not to do A, which is to say that R- in fact outweighs R+. But if you do A anyway, you thereby demonstrate that you regard R+ as outweighing R- (assuming you are rational). In these conditions, it follows that you will be *insufficiently repelled* by the harm in the sense that the weight attached to the harm as a reason not to do A was *insufficient to get you to refrain from doing A*. Furthermore, this does not turn insufficient repulsion into an empty notion because it means nothing more than not doing the right thing.[66] After all, one might be sufficiently repelled by a harm in the comparative sense but still fail to avoid causing it due to execution failure when trying to act as planned.[67] Moreover, even if insufficient repulsion and incorrect action are co-extensive when one is fully rational and there is no execution failure, the two notions would still have different *intensions* because one concerns action and the other concerns motivation. Thus, the notion of insufficient repulsion does not just collapse into that of incorrect action.

Now return to the putative counter-example from above. We can see it poses no threat to premise 3a) when this comparative notion of insufficient repulsion is employed. Despite being substantially repelled by the harm (5 units), the actor was even more attracted to it (6 units) in virtue of the benefits she believed it would lead to. Nonetheless, she was still insufficiently repelled by the harm in the sense that her awareness of it *should* have given her a sufficient

[66] Thanks to an anonymous referee for pressing me on this point.

[67] For example, I might be appropriately motivated to dive into the freezing water to save a drowning child, but as I am about to dive, a prior trauma kicks in and paralyzes me, thus preventing me from acting as I am motivated to do. Accordingly, I would fail to do the right thing even though I am sufficiently repelled by the harm I would cause by not saving the child. Thus, insufficient repulsion is not simply *defined* as failing to do the right thing.

motivating reason to refrain from doing the harmful act—*i.e*. this was the job her motivational apparatus was called on to do—but in fact the only motivating reason the harm gave her against doing the action was insufficient to get her to avoid imposing the harm. Thus, this example is no threat to premise 3a)'s claim that when one is committed to bringing about an unjustified harm, one will be insufficiently repelled by it. (As we'll see in considering premise 4), the same holds for those who merely foresee that they will cause unjustified harm.)

Premise 3b)

When it comes to acting with the *commitment* to harming, however, this is not the end of the story. In addition to being insufficiently repelled, 3b) claims that when you act with some degree of commitment to causing the harm, this shows that you feel some affirmative motivational pressure, under the circumstances of the case, towards promoting the harm. That is, you feel some affirmative inclination to take steps, or make adjustments to your conduct, so as to *make the harm more likely* (*i.e*. more likely than it would be just given A without those steps or adjustments).[68] Another way to put it would be to say that your conduct is *guided* by your commitment to the harm. Your commitment pushes you, as it were, towards more effective means to bringing about the harm, were any available, under the circumstances of the case. Moreover, even if no such steps are obviously available that would promote the harm more than your actual action A does, you will—in virtue of your commitment to the harm—still be on the *lookout* for such steps. After all, were they to reveal themselves, you would feel some motivational pressure to avail yourself of them. So even if none turns up, you still are guided by your commitment in the sense that you *seek* such more effective ways to bring about the harm.[69]

[68] Alternatively, we might say that in virtue of your commitment to the harm, you would be *rationally criticizable* for failing to feel any motivational pressure towards promoting the harm. Either claim would do.

[69] Hills and Bratman make similar claims. Hills thinks that intending a state of affairs, S, entails a commitment to S in the sense that one "chooses actions on the basis of their contribution" to bringing about S and "monitors" one's success at bringing it about. *See* Hills, *supra* note 54 at 135-36. It is this claim about *monitoring* that resembles my

This, in turn, shows that you take there to be reasons that count in favor of promoting the harm. For unless you took it that there was something that counted in favor of the harm itself (either intrinsically or because it is a means to something else you want), you would not be inclined to seek and take steps that appear to promote it. Thus, when you are committed to the harm, we know that you are drawn to the prospect of the harm and thus take its occurrence—or more precisely, the facts that seem to count in favor of the harm—as being included in the set of considerations that count in favor of A, *viz*. R+.

Thus, assuming both A and the harm to which you are committed are *unjustified*, you will—in addition to being insufficiently repelled by the harm—display this second distinctive fault. This is the fault of being affirmatively attracted to the prospect of the harm *more than you should.* Your commitment to the harm demonstrates that you attach some positive weight to considerations that either do not count in favor of the harm at all (and thus do not count in favor of A), or that only count in favor of the harm (and thus A) *less strongly* than you take them to, given that the harm (and A) are unjustified. After all, by doing A with a commitment to the harm, you demonstrate that you take there to be sufficient reason to cause the harm, but in fact, by hypothesis, there is not: the harm is unjustified (as is A). Thus, any reasons there might be in favor of the harm actually support it less strongly than you take them to. Hence, you are more attracted to the harm than you should be.[70]

To this, one might object as follows. Suppose your act is itself on-balance unjustified, not because the *harm* you mean for it to cause is unjustified (suppose the harm *is* justified), but rather because of some *other* feature of the act that makes it on-balance unjustified. Perhaps the act is on-balance unjustified because it is a *promise breaking*, and although you also intend the

claim about being on the *lookout* for ways to better promote the intended harm. Likewise, Bratman claims that "[i]n the normal case, one [who intends a given effect] is prepared to make adjustments in what one is doing in response to indications of one's success or failure in promoting" that effect. Bratman, *supra* note 54 at 141.

[70] As with insufficient repulsion, over-attraction must also be understood comparatively.

harm you know it will cause, this harm itself is justified on independent self-defense grounds. For instance, suppose your act will prevent someone from unjustly punching you in the face although it will harm your attacker, and your act also happens to amount to breaking a solemn promise. Thus, suppose the action is on-balance unjustified not because of the harm it will cause, but because of the overriding badness of breaching a solemn oath. Wouldn't this case—call it *Treacherous Self-Defense*—be one in which you do *not* obviously overvalue the reasons in favor of promoting *the harm itself*?

The answer is yes, but it doesn't matter. Granted, this would be a case where, despite doing an action that is on-balance unjustified, you do not attach greater weight to the reasons in favor of *the harm* than you ought to. Thus, I grant that in this case, you do not display the "second distinctive fault" mentioned in 3b). Nonetheless, this is not a problem for premise 3), because 3) is restricted to cases in which *the harm to which you are committed is itself unjustified*. 3b) claims only that if you do an action A with the commitment that A bring about some harm H, then provided *neither A nor H* is justified, *then* you display the second distinctive fault of attaching greater weight to the reasons in favor of promoting H than you ought to. Accordingly, Treacherous Self-Defense does not threaten the claim on which the Two Strikes Argument relies.

Premise 4)

Next consider premise 4). Unlike intending a harm, when you merely foresee or know that a given harm will result from your action A, but you are in no way *committed* to it, then you do not necessarily take there to be any reason in favor of promoting the harm. Rather, the only fault you display is that you are insufficiently repelled by the harm you know A will cause. You do not take there to be any reason to ensure the harm comes to pass, or to make it any more likely than it already is given A. You are not on the lookout for additional steps or adjustments to your conduct that might further raise the probability of the occurrence of the harm. Instead, your only

fault is that you are insufficiently repelled by the harm in the sense that you take the reasons in favor of A, namely R+, to outweigh, the reasons against A, namely R-, even though in reality R- outweighs R+ (and you should have been motivated accordingly).

Thus, what seems to be going on in Arson 1 is that Alan takes the receipt of $5000 to be a reason in favor of lighting the fire, and he is morally criticizable because he failed to attach sufficient weight to the obvious reason against lighting the fire: namely, that it clearly will result in Victor's death. Either Alan attached no weight to it at all, or he attached far too little weight to it (more precisely, too little weight compared to the fairly minor benefit of getting $5000[71]). So in Arson 1, Alan only failed to be sufficiently repelled by the harm he knew his act would cause, but he did not feel any affirmative attraction to, or take there to be positive reasons in favor of, promoting the harm *itself*. Thus, he displays only one of the distinctive faults Bobby displayed.

Now, the obvious objection to line 4) is that merely knowing of or foreseeing a particular harm does not *preclude* one's seeing it as a good thing that the harm comes to pass, even when one is not committed to it. For example, imagine a version of Arson 1—call it *Happy Side Effect*—where Alan's only motivating reason for lighting the fire is that he wants the $5000 (just as before). But, in addition, suppose now that he also happens to hate Victor. Accordingly, Alan regards Victor's likely death in the fire as a happy side effect of his starting the fire. Doesn't this show that in some cases of merely foreseen harm the actor really *will* feel an affirmative pull towards the harm, and not simply be insufficiently repelled by it, as I have claimed? That would undermine the explanation I am trying to offer of why merely foreseen harm is (all else equal) less culpable than harm to which you are committed.

The answer to this worry can be seen by focusing on what it is for some bit of conduct to *manifest* insufficient regard. Granted, it is possible to merely foresee harm while at the same time

[71] As noted, what matters to culpability is how strong the motivational pressure one feels against the action is *compared to* the motivational pressure one feels in favor of it.

regarding it as a fortunate side-effect of your conduct. Nonetheless, in cases of merely foreseen harm, to which you are not committed, any positive draw you feel towards the harm will not be doing any work in explaining your conduct and thus will not be *manifested in it*. Let me explain.

There is a crucial difference between bad character and culpable conduct, and it is only the latter that we are interested in here. After all, it is a fundamental principle of the criminal law that we do not punish merely for bad attitudes or character traits one might possess, but only for conduct that manifests them. As Ken Simons explains:

> 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. (…) We are similarly, and properly, reluctant to impose punishment on a person simply for [attitudes or characteristics] unless and until [they] are *expressed in action*."[72]

It is for this reason that when Bill drives off with the intention to kill his uncle but carelessly hits and kills a pedestrian who turns out to be his uncle, we do not say Bill is guilty of murder.[73] Granted, we know that his *character* is as deplorable as a murderer's, given that he *would have* murdered his uncle had he gotten the chance. But the insufficient regard inherent in murder is not manifested in *the act at issue here*—Bill's careless driving. Recall the principle of lenity mentioned earlier. It holds that an act only manifests the *minimum* amount of insufficient regard it is necessary to postulate in order to explain why the actor behaved as she did under the circumstances.[74] But in order to explain why Bill drove carelessly in this case, we do not need to posit anywhere near as much insufficient regard for others as is characteristic of murder (*i.e*. an unjustified intentional killing). Accordingly, under the principle of lenity, we may only take

[72] Kenneth 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, 233-34 (2002) (emphasis added).
[73] *Id*. at 232.
[74] Gideon Yaffe defends a version of this principle of lenity in *The Point of Mens Rea: The Case of Willful Ignorance* (draft). Thomas Aquinas endorses a similar principle: "unless we have evident indications of a person's wickedness, we ought to deem him good, by interpreting for the best whatever is doubtful about him." *Summa Theologica* II-II 60, 4. Aquinas argues that "from the very fact that a man thinks ill of another without sufficient cause, he injures and despises him." *Id*. Thus, he concludes, we ought to apply a principle of lenity when attributing blame to actors for their conduct.

Bill's conduct to manifest as much insufficient regard as the lowest amount needed to explain his conduct under the circumstances. Otherwise we would be blaming him just for his bad character, *i.e*. based on attitudes that do no work in explaining his conduct.

In general, the thought here is to take the *least bad configuration* of mental states that would be sufficient to produce the action under the circumstances in question, and then suppose that one's actual action only *manifests* the insufficient regard associated with *those* mental states. Any additional deplorable motives or mental states you might have acted with—while they might impact the badness of your character—are not going to count as *manifested in your action*. That, in turn, is because they do not do any work in explaining your action under the circumstances. While mental states that do not help explain your conduct may bear on the condition of your character, they do not affect your culpability for that conduct itself.

This point carries over to *Happy Side Effect*. There, Alan's dislike of Victor and gladness over the prospect of his death shows that Alan has a deplorable character. But Alan's dislike of Victor nonetheless is not *manifested* in his lighting the building on fire. Why? Because we do not *need* to postulate that he feels this motivational pull towards the harm in order to explain why he lit the building on fire in this case. In order to explain it, all we need to appeal to is i) the fact that he took receiving $5000 to count in favor of the action, and ii) that this motivational pull towards the action failed to be outweighed by any motivational repulsion in virtue of the fact that lighting the fire would kill Victor. But to explain Alan's conduct, we do *not* need also to point to iii) the fact that he is happy to see Victor get killed. Rather, i) and ii) are sufficient to get an actor to perform the action under the circumstances, and iii) is superfluous. Thus, while it's true that Alan's character is shown to be worse by his happiness about Victor's death, the culpability of Alan's conduct only depends on how much insufficient regard it *manifests—i.e*. on i) and ii).

One might object that I have merely *asserted* that Alan's dislike of Victor in *Happy Side Effect* is not manifested in his act of lighting the fire, since this dislike is not needed to explain why Alan acted as he did (*i.e*. is no part of what motivated him to do so). How can we be *sure* this is so? Here is an argument to secure the point (and thus premise 4)).

The claim to be shown is that Alan's dislike of Victor in *Happy Side Effect* is no part of what actually motivated him to light the building on fire, and thus is not manifested in his action. Now suppose for *reductio* that his dislike of Victor really *was* part of what actually motivated him to behave as he did, *i.e*. actually did exert some motivational pull on him, which we have to appeal to in order to explain his conduct. If it *were* part of what actually motivated Alan, then we can expect that if there had been several different ways of lighting the fire, some of which appeared to make Victor's death more likely than others, then Alan would have felt some motivational pressure to opt for one of the ones that seem to make Victor's death more likely. However, that would entail that Alan was *in fact somewhat committed* to Victor's death in the sense defined above. However, by hypothesis, Alan is *not committed at all* to Victor's death. That was part of the setup of *Happy Side Effect*. It was supposed to be a case of merely foreseen harm, to which the actor was not committed, but where he still feels some motivational pull towards the harm (*i.e*. takes it that there are some reasons in favor of promoting the harm). But this leads to contradiction. If we suppose that Alan's lighting the fire was actually partially motivated by his dislike of Victor, then he would be somewhat committed to the death, even though Alan was stipulated *not* to be committed to it at all. Accordingly, we see that an outcome or state of affairs cannot actually exert motivational pull on you without it also being the case that you are somewhat committed to bringing it about. Thus, if you do an action while *merely foreseeing* that it will cause harm, then it cannot be the case that the harm actually exerted any motivational pull

on you in a way that is necessary to explain your action. For if it did, then you would in fact be somewhat committed to the harm, and so you would no longer qualify as *merely* foreseeing it.

Hence, it must be the case that in *Happy Side Effect*, Alan's dislike of Victor is no part of what actually moved him to light the fire, and thus does no work in explaining why he behaved as he did. At best, it is an inert, free-floating sentiment that is not manifested in his lighting the fire. As a result, Alan's dislike of Victor in *Happy Side Effect* forms no part of the basis for Alan's culpability for his conduct (though it clearly still bears on the quality of his character).

Accordingly, I take it that the point needed for premise 4) has been established. That is, when you merely *foresee* that harm H will result from your action, but are not committed to it, you either do not take there to be any positive reason in favor of promoting the harm (*i.e*. experience any motivational pull towards it) *or* if you have some kind of positive attitude towards the harm (as in *Happy Side Effect*), then this fact still is not *manifested* in your action in a way that bears on your culpability for that action. Instead, in cases of merely knowing or foreseen harm, the actor's conduct only manifests an insufficient repulsion from the harm.[75]

---

[75] Steve Finlay raised the following interesting objection to Premise 4). Suppose in a modified version of *Happy Side Effect*, Alan would not have started the fire *except* for the fact that he noticed it would kill specifically Victor, Alan's sworn enemy, and Alan sees this as a good thing. Although an oversimplification, suppose Alan initially felt 10 units of attraction to lighting the fire (in virtue of the money it would get him), but he also felt 11 units of repulsion to it in virtue of the harm it would cause. But his realization that the fire would kill specifically Victor *weakens* Alan's repulsion to the harm from 11 down to only 9 units. Thus, Alan is now able to light the fire and he does. Here, Alan sees Victor's death as a good thing and it would be genuinely *manifested* in his conduct (since it is a but-for cause of that conduct), but Alan still would not be *committed* to Victor's death to any degree. After all, the fact that the victim is Victor only *neutralized* some of Alan's repulsion to the harm. So we seem to have a counter-example to Premise 4). After all, even though Alan here is supposedly not committed to the death, it seems he still sees positive reasons in favor of the harm itself. Therefore, he *would* display the second "strike" associated with intended harm.

Nonetheless, it seems doubtful that normal human psychology allows it to be simultaneously true that consideration of a harm weakens one's motivational repulsion to it, *and* that one has no commitment whatsoever to the harm. For this case to be a counterexample to Premise 4), three things must be true: i) the actor has to merely foresee the harm and not be committed to it; ii) he has to do the action in part *because* it will be harmful (after all, his regarding the harm as good needs to be manifested in his conduct); and iii) he must take there to be positive reasons in favor of the harm itself. But there is tension between i) and iii), at least assuming the actor's psychology is realistic. If an actor saw reasons in favor of the harm itself (here, the fact that the victim is an enemy), and this is manifested in his conduct, then he plausibly would be *at least somewhat committed to the harm itself*. Commitment, in my sense, involves being inclined to act in ways that make the harm more likely, and being on the lookout for ways to do so. If one really saw reasons in favor of the harm itself, and if this really is sufficient to weaken the

Putting the pieces together

Now we are in a position to see how the conclusion—line 5)—follows from premises 1)-4). When you do A while merely foreseeing that it will cause harm H, then assuming neither A nor H is justified, you are not sufficiently motivationally repelled by H to prevent you from doing A for whatever reasons you thought counted in favor of it. This manifests some degree of insufficient regard for morally relevant interests. By contrast, when you are *committed* to H, such that you take there to be positive reasons in favor of promoting it (*i.e*. when you feel some motivational pull towards steps that would further increase the probability of H occurring), this actually involves *two faults*. First, you failed to be sufficiently motivationally repelled by H—that is, in doing A, you failed to regard the harm as *as* strong a reason against A as you should have. Thus, your action manifests insufficient regard in the same way as what was the case with merely foreseen harm. But *in addition*, the second fault is that you also wrongly took there to be reasons in favor of promoting H (at least assuming H and A are unjustified). That is, assuming H is not supported by the applicable normative reasons, you *either* felt some motivational pull towards it when you should not have, *or* you felt more motivational pull towards H than was warranted by the actual normative reasons in favor of H. This second fault demonstrates an *additional* amount of insufficient regard for morally relevant interests. Thus, acting from a commitment to the harm manifests insufficient regard and is culpable in *two ways*, while acting while merely foreseeing the harm manifests insufficient regard and is culpable in just *one way*. Therefore, there is one respect in which the former is more culpable than the latter. Hence, DDE_{NACR} holds assuming A and H are unjustified.

---

repulsion one initially felt towards the harm, then most normal people would *also* feel some inclination to do things that make the harm more likely and be on the lookout for such harm-promoting steps. That is, they would indeed be at least somewhat *committed* to the harm. Of course, it might be logically possible to have the sort of psychology that undermines Premise 4). But it seems doubtful that normal humans are actually like this. Such a psychology would seem strangely disjointed and compartmentalized.

4. Objections

We now have an explanation for why the action in Arson 2 is more culpable than that in Arson 1. I will close by discussing three objections to the Two Strikes Argument. First, one might worry that the argument contains an implicit premise: namely, that the two faults the argument identifies add up to greater culpability than just the one fault by itself. This premise might be questioned.[76] After all, sometimes an action can be faulty in several ways, which do not all contribute distinct amounts of culpability to it. Suppose someone criticizes an action of mine because it not only is an instance of running a red light at a busy intersection, but also because it is a dangerous traffic violation. These two faults seem to collapse into one another, such that the latter fault does not add any more culpability beyond what is already conferred by the first. On the other hand, an action might display genuinely distinct faults that do not collapse into each other. If I promised to scrupulously obey the traffic laws, then my running the red light would not only be a dangerous traffic violation, but it would be a promise-breaking as well. These two faults are distinct, and therefore do seem to confer distinct amounts of culpability. How do we know if the two faults displayed by intentional harming are distinct, rather than collapsible?

The two faults displayed by intentional harming—failing to be sufficiently repelled by the harm and being overly attracted to it—are distinct, I claim, because they are *failures of different motivational mechanisms*. On the model presupposed by my prior discussion (especially of 3a), it makes sense to think of our motivational machinery as involving both an attraction mechanism and a repulsion mechanism, which have distinct jobs (*i.e*. different criteria for success).[77]

[76] Thanks to an anonymous reviewer for this objection.

[77] Some empirical also work suggests that our actions are governed by both an attraction mechanism and a distinct aversion mechanism. *See, e.g.,* C. Carver & T. White, *Behavioral Inhibition, Behavioral Activation and Affective Responses to Impending Reward and Punishment*, 67 J. PERSONALITY & SOCIAL PSYCHOLOGY 319 (1994) (discussing Jeffrey Gray's influential work on the behavioral activation and inhibition systems).

Begin with the attraction mechanism. One of its jobs is to get you to be drawn to the things you should be motivated to bring about, while another aim—more important here—is to *avoid* generating an attraction to things you shouldn't be motivated to bring about. Thus, when you are attracted to an unjustified harm, which necessarily is something you should not be attracted to, your attraction mechanism misfires. That is one kind of motivational failure. However, it seems we also have a failsafe mechanism for just such an occasion: namely, the repulsion mechanism. (As any engineer knows, some redundancy is an important part of successful systems design.) Part of the job of the repulsion mechanism is to repel you from bad states of affairs enough to get you not to bring them about (at least when there is no justification for doing so). Thus, if you perform an action intending to bring about an unjustified harm (and you are rational), we can infer that this failsafe itself has failed. You were not repelled by the harm sufficiently to get you to refrain from the act in question. The amount of repulsion to harming that is required in a given case depends in part on how much overt attraction you have towards the harm. If your repulsion mechanism does not produce the amount of repulsion needed to get you to refrain from the harmful action, it has failed to do its job, and this is a distinct kind of failure from the misfiring of the attraction mechanism.

Thus, when you act intending to cause harm, there appear to be two different mechanisms that fail—one involving attraction, the other involving repulsion. The mechanisms are fairly seen as distinct given that they have different jobs. Thus, their failings are distinct, and do not appear to collapse. This is the reason to think the culpability of both faults identified by the Two Strikes Argument add up to more culpability than either one alone.

A second objection to the Two Strikes Argument is that there might seem to be no guarantee that those who intentionally cause harm are more culpable than similarly situated actors who cause analogous harms only knowingly. Consider the following variation of Arson 1 & 2.

Suppose Charlie (just like Alan) lights the fire for $5000 despite knowing that it will kill Victor. Moreover, Dan (just like Bobby) lights the fire intending to kill Victor in order to get the $5000. But now suppose that Dan regards Victor's death as deeply troubling and sees it as a strong reason not to light the fire. Although he proceeds to set fire to the building, he does so with strong reservations. (Perhaps he thinks he *really* needs the $5000.) By contrast, Charlie, who merely foresees that the fire will kill Victor, feels no reservations whatsoever about Victor's death. Charlie couldn't care less that Victor will die. Given that Charlie thus seems extremely callous, while Dan seems somewhat less callous, one might conclude that intentionally harming is not always more culpable than analogous acts of causing harm knowingly.[78]

This objection admits of two answers. First, my aim in this paper is only to argue that *there is one respect* in which intending harm is more culpable than bringing about a similar harm only knowingly, and this holds even for Charlie and Dan. After all, while Charlie's and Dan's respective actions share one culpability enhancing feature (*i.e*. neither was sufficiently repelled by the harm to get them to refrain from starting the fire), Dan's action also manifests insufficient regard, and thus is culpable, in a second way. Because Dan is committed to the harm (and thus is on the lookout for steps he might take to promote its occurrence, *etc*.), we know he took there to be positive reasons in favor of the harm itself that he regarded as sufficient grounds for bringing it about, even though by hypothesis there were no reasons sufficient to justify the harm. By contrast, Charlie's conduct does not display this second fault, since he merely caused the harm knowingly and was not committed to it. Accordingly, Dan's conduct has one culpability-enhancing feature that Charlie's lacks. Thus, even if Charlie is on-balance more culpable than Dan, this is compatible with the conclusion I am arguing for here.

[78] Thanks to Jon Quong for pressing me on cases of this kind.

The second answer is that the case is under-described because it does not mention the benefits Charlie and Dan seek. When the case is plausibly filled in, we see that it either is a case where not all else is equal (and thus not a problem) or one where Charlie's conduct is not worse than Dan's (even if Charlie's character is worse). On the one hand, the case might be filled in such that the reason for Dan's reservations and Charlie's indifference is that they do not see the same reasons in favor of Victor's death. On this reading, Dan thinks there are powerful reasons that support the death (*e.g*. that it will allow him to afford something of importance like medicine for his child, *etc*.), while Charlie does not; he wants the money for personal amusement and is just indifferent to Victor's fate. In that case, the actors would be seeking benefits of different magnitudes and so all else would not be equal. Thus, it would not be problematic if Charlie *on these grounds* is more culpable than Dan. For all else to be equal, also the benefits sought by the two actors would have to be of comparable magnitudes.[79]

On the other hand, we might flesh out the case so that all else *is* equal. But then Charlie's *conduct* would not be more culpable than Dan's (even if Charlie's character might still be worse). Thus, suppose that the benefits sought by both are the same—*e.g.* $5000 just for personal amusement. This would show that both Charlie and Dan are so unconcerned with the value of Victor's life that it only takes the promise of $5000 to get them to be willing to accept his death. Granted, Charlie was said to be a callous fellow who is *even less* concerned with Victor's life than Dan was. This means Charlie would have been willing to tolerate Victor's death for *even less* money than Dan. But that is only a fact about how Charlie would have been willing to act under non-actual circumstances, and so his actual conduct does not manifest this fact about what he'd be willing to do. All it shows is that Charlie's *character* is worse than Dan's. But it doesn't directly bear on the culpability of his actual *conduct*. Accordingly, when all else is held equal,

[79] *Cf. supra* note 33.

Charlie's conduct seems no more culpable than Dan's, even if Charlie's character might still seem worse. In fact, Dan's actual conduct is arguably worse because it involves an overt commitment to Victor's death, while Charlie's does not. Thus, whichever way the case is read, it would pose no problem for the conclusion of the Two Strikes Argument.

The third objection is more troubling. Whether one merely foresees a harm or intends it often will only depend on contingent circumstances.[80] For example, it is just a matter of luck that in Arson 1, Tony offered Alan $5000 to burn down the building, while in Arson 2, Tony offered Bobby the same amount of money to kill Victor. Thus, Victor's death is not necessary as a means to Alan's getting paid, but it is necessary as a means to Bobby's getting paid. Why should such contingent facts about what happens to be necessary as a means ground a difference in the culpability of the two actors in question? After all, both Alan and Bobby seem to be to blame, at bottom, because they attach insufficient value to human life.

Although I feel the force of this objection, let me offer two replies. First, on any version of DDE, the instrumental relationship a given harm happens to have to one's ends will affect the moral status of one's conduct. Thus, it seems one cannot preserve any version of DDE without accepting that some moral luck of this kind exists. Second, and more importantly, it seems that contingent circumstances can *generally* make a difference to culpability—not just in the DDE context. Sometimes we behave worse than others simply because we are unlucky enough to be presented with the opportunity to do a greater evil (*e.g.* if we happen to live in violent or oppressive societies). If we take the opportunity presented, we will have done something worse than those who were fortunate enough to not be offered the chance to do a comparable evil. The same seems true in Arson 1 & 2. Bobby was presented with the opportunity to perform a worse action then Alan was, and he took it. Although Alan in this sense lucked out, it is still seems that

[80] Bratman raises a similar concern. *See* Bratman, *supra* note 54 at 161. Thanks also to Jon Quong for pressing me on this point.

what Bobby did (intentionally kill without justification) is more culpable than what Alan did (knowingly cause death without justification).

5. Conclusion: The Closeness Problem Redux

I have defended an argument in favor of a restricted, non-absolute, culpability-based version of the doctrine of double effect, which I dubbed DDE_{NACR}. This provides a rationale for the top tier of the criminal law's culpability hierarchy, and explains why it makes sense to sometimes attach heavier sanctions to crimes (like treason) that require *purposefully*, not just knowingly, bringing about a bad state of affairs. My argument, as an agent-centered rationale, promises to do a better job than the means rationale of providing a complete normative foundation for the way in which the criminal law actually encodes DDE.[81]

Of course, this still leaves the closeness problem. I have not tried to fully solve it. Admittedly, DDE_{NACR} will not be completely defensible until we have a principled way to distinguish intended and merely foreseen effects. However, at the very least my agent-centered rationale for DDE_{NACR} is no worse off on this score than the means rationale, which itself requires that we can distinguish between using (or intending to use) others as means and merely affecting people more indirectly.

What is more, my explanation for why purposeful misconduct is more culpable than merely knowing misconduct (assuming there are no relevant justifications and all else is equal) also provides some practical guidance in distinguishing the two categories from one another. I argued

[81] Another question I have not addressed is why the criminal law often *declines* to single out the purposeful version of some type of misconduct for harsher treatment than the merely knowing version of that conduct. (Tort law, by contrast, collapses this distinction. *See* Restatement (Second) of Torts § 8A (1965).) Perhaps part of the explanation rests on the assumption that once some threshold of culpability is passed that triggers the harshest penalties the law has available—as is the case with murder—it's no longer necessary to distinguish purpose from knowledge. Thus, knowingly killing might surpass the threshold of culpability needed to merit the harshest penalties. But then we need to explain why treason, which also can trigger severe penalties, *does* distinguish purpose from knowledge. Maybe since treason can be committed without anything as bad as killing, only the purposeful version of this misconduct passes the threshold needed to trigger the harshest penalties. However, more work is needed here.

that acting with purpose to bring about a bad or wrongful state of affairs is *pro tanto* more culpable than doing so merely knowingly because the former involves a certain type of *commitment*, which the latter lacks. Accordingly, to distinguish purposeful from merely knowing misconduct, what we need to look for is evidence of this type of commitment. Thus, evidence will be probative of whether the defendant intended a given harm, as opposed to merely foreseeing it, when this evidence tends to show that the defendant was motivated to *promote* the occurrence of this state of affairs under the circumstances. Did the defendant commit the crime in a way that made the occurrence of the bad state of affairs more likely than other ways she easily could have committed the crime instead? Was she on the lookout for ways to increase the chances that the bad state of affairs would occur? Evidence indicating an affirmative answer to such questions is evidence that the defendant was committed to the bad state of affairs in the relevant sense. Accordingly, this is one crucial type of evidence we need to look for to decide if a defendant manifested both kinds of fault that make intending harm worse than foreseeing it.

In this way, my argument has the added benefit of giving some practical advice for how to decide whether a defendant should be treated as a purposeful or a merely knowing actor. This, in turn, helps make progress in dealing with the closeness problem as it is likely to come up in legal contexts. Because of this and the other advantages of my defense of DDE_{NACR}, we have reason to be optimistic that the doctrine of double effect as actually employed in the criminal law can be placed on a sound normative foundation.