Published in *Review of Philosophy and Psychology*, 5: 505-525 (2014).
Please cite the published version: http://link.springer.com/article/10.1007/s13164-014-0192-5

Distinguishing between three versions of The Doctrine of Double Effect hypothesis in moral psychology

Simon Fitzpatrick
Department of Philosophy
John Carroll University
sfitzpatrick@jcu.edu

Abstract

Based on the results of empirical studies of folk moral judgment, several researchers have claimed that something like the famous *Doctrine of Double Effect* (DDE) may be a fundamental, albeit unconscious, component of human moral psychology. Proponents of this psychological DDE hypothesis have, however, said surprisingly little about how the distinction at the heart of standard formulations of the principle—the distinction between intended and merely foreseen consequences—might be cognised when we make moral judgments about people's actions. I first highlight the problem of precisely formulating the distinction between intended and foreseen consequences and its implications for interpreting the empirical data on folk moral judgment. I then distinguish between three different approaches to this problem that have been taken by proponents of the DDE in normative ethics: so-called "closeness" accounts, accounts that employ what has come to be known as a "strict" notion of intention, and Warren Quinn's recasting of the DDE in terms of the distinction between "direct" and "indirect agency". I show that when taken as claims about moral psychology, these different accounts entail quite different empirical predictions about what people's moral judgments should be in particular cases. Based on the current empirical data, I argue that a version of Quinn's formulation of the DDE is the most empirically plausible, and that adopting such a formulation helps to diffuse much (though not all) of the recent empirical criticism of the DDE hypothesis.

Keywords: Moral psychology, Doctrine of Double Effect, Trolley problem, Intention

1. Introduction

The Doctrine of Double Effect (DDE) is a principle of normative ethical justification, which, traditionally formulated, draws a moral distinction between causing a morally grave harm to another person as a foreseen but unintended side effect of pursuing a good end, and causing such a harm as an end or as an intended means to achieving one's end. When certain other conditions are met, the claim is that actions causing harmful

consequences in service of a good end may be permissible when the harmful consequences are merely foreseen and not directly intended or intended as a means to one's end.[1]

The DDE has been, and remains, highly controversial in normative ethics (see, e.g., the papers in Woodward, 2001). However, in recent years there has been increasing interest in the DDE from researchers in cognitive science. Though this research may have normative implications (see, e.g., Mikhail, 2011; Greene, 2013), the core issue here is not whether the DDE is a sound normative principle, but whether it is a fundamental, albeit unconscious, component of human moral psychology, and thus captures part of the implicit logic behind the moral evaluations of ordinary folk. In particular, a tacit psychological version of the DDE has been thought to explain various patterns that have been discovered in folk moral intuition—that is to say, in the immediate moral judgments that people make in response to moral stimuli. For example, in numerous studies employing variants of the famous "trolley problem", which has been widely discussed in normative ethics (e.g., Foot, 1967; Thomson, 1985), participants have tended to say that it would be morally permissible for a hypothetical bystander to flick a switch to divert a runaway trolley that is about the kill five people onto a side track at the cost of killing one person stuck on the side track (Switch case), but that it would *not* be morally permissible for a bystander to push a large man off a bridge in front of the trolley to stop it (Footbridge case), even if this would bring about the same good result of saving the five at the same cost of killing one person (e.g., Mikhail, 2000, 2011; Cushman et al., 2006; Hauser et al., 2007; Greene, et al., 2009; Arbarbanell and Hauser, 2010). This difference in moral intuition, which has been found in participant pools ranging over a variety of ages and cultural and ethic backgrounds, has been claimed as highly suggestive evidence that the DDE may be ingrained into folk moral cognition: it seems that when we consider these and other similar moral dilemmas, we intuitively distinguish between the moral

[1] The DDE is usually traced back to Aquinas' discussion of self-defence in *Summa Theologica*, though contemporary discussions arguably owe more to nineteenth-century formulations (see Connell and Kaczor, 2013). Other conditions of the principle typically include a proportionality condition (that the moral benefits brought about by the action outweigh its harmful effects), and the condition that there be no better alternatives (e.g., a way to achieve the good end without causing any harm). Proponents of the principle also tend to emphasise that it should not be taken to be a complete moral theory, but only a component of one. Thus, the verdict of the DDE with respect to a particular act ought not be taken as the last word on the matter.

acceptability of an action in service of a good end that causes harm to a person as a *foreseen side effect* (Switch case), and an action that involves causing harm *intentionally* as a means to a similarly good end (Footbridge case). Thus, the cognitive processes that produce these intuitive moral judgments may (unbeknownst to us) actually be applying to the relevant situations a tacit psychological principle directly analogous to the DDE discussed in normative ethics.[2] Moreover, researchers such as John Mikhail (2000, 2011) have suggested that since the DDE is a relatively complex moral principle that would be difficult to infer from the moral stimuli that children are typically exposed to during development, it may in fact be an *innate* moral principle, perhaps part of an innate "universal moral grammar".

These claims (particularly the last one) have attracted much attention and criticism (see, e.g., Prinz, 2008; Sterelny, 2010).[3] My goal in this paper is to explore what has so far been a neglected issue in discussions of the DDE as a candidate component of human

[2] These studies (especially those reported by Mikhail, 2000, 2011) have also found effects corresponding to other components of the DDE, such as proportionality and no better alternative conditions. It should also be noted that the apparent effect of intentions on moral judgment seems hold even in the absence of the use of personal force and direct physical contact, both potentially confounding factors in the standard Footbridge case (Mikhail, 2000, 2011; Cushman et al., 2006; Hauser, et al., 2007; Arbarbanell and Hauser, 2010; Huebner, et al., 2011). Though the pattern is not entirely uniform (see, e.g., Greene et al., 2009; Greene, 2013), most studies find that participants still tend to judge it impermissible to use the large man to stop the trolley from killing the five in "impersonal" versions of the Footbridge case where the agent does not have to push the large man off the bridge or use any personal force to move the large man in front of the trolley (e.g., where the agent instead pushes a button to drop the large man in front of the trolley). Such impersonal versions of this scenario do receive relatively higher permissibility ratings, suggesting that the use of personal force does have an interacting effect with intention (see Greene et al., 2009). Nonetheless, such impersonal versions of the Footbridge case still tend to receive significantly lower permissibility ratings than those for the Switch case.

[3] One criticism of this research concerns the putative artificiality of the trolley-style scenarios that have been used, leading to the worry that the results of these studies may not get at any real underlying psychological principles, but only reflect how people respond to unusual hypothetical cases. However, defenders of this research usually respond that while the scenarios are artificial and unrealistic they do enable us to control for the influence of other factors on moral judgment. Moreover, while concerns have also been raised about the extent to which participants' judgments may be influenced by various framing and order effects—and hence might be less systematic than they appear (e.g., Schwitzgebel and Cushman, 2012; see also Liao et al., 2012; Wiegmann et al., 2012; Wiegmann and Waldmann, 2014)—the patterns in moral judgment revealed by these studies do seem to be quite robust and cannot be fully explained away by either the artificiality of the scenarios (one might expect participants' judgments to be all over the place if that were so) or by framing and order effects. Another concern is that the results usually reveal a significant minority opinion as well as a majority one—for instance, it is almost always the case that a number of participants judge it permissible to push the large man in the Footbridge case. As result, one might be sceptical about using such studies to draw general conclusions about human moral psychology. This is, however, a pervasive phenomenon in psychological research, and thus not especially problematic for making general claims about moral psychology—though, of course, it is important to consider how such diversity in responses is best explained.

moral psychology. Proponents of this psychological DDE hypothesis have said surprisingly little about *how* the distinction at the heart of standard formulations of the principle—the distinction between intended and merely foreseen consequences—might be cognised when we make moral judgments about people's actions: i.e., under what sorts of circumstances do the hypothetical underlying cognitive processes mentally represent the harmful consequences of an agent's action as intended (e.g., as a means to a particular end) versus merely foreseen when applying the principle?[4] The claim has just been that empirical studies, such as the trolley studies just mentioned, seem to show that we do cognise such a distinction, and hence that a tacit psychological version of the DDE offers a plausible explanation for why we make the intuitive moral judgments that we do in response to these sorts of cases.[5]

There is, however, a worry lurking here that is closely linked to one of the main objections to the DDE in normative ethics, which stems from the concern that the distinction between intended and merely foreseen consequences is left far too *vague* to do any genuine explanatory work with respect to our moral intuitions. Indeed, some critics have argued that no principled account of the intend/foresee distinction can in fact properly explain the moral intuitions that the DDE has been thought to capture and ultimately justify (e.g., Davis, 1984; Bennett, 1995). As soon as one tries to formulate this distinction more precisely, so that the DDE clearly rules against the actions it is

[4] There is, it should be noted, a growing literature in experimental philosophy on the folk concept of intention and its connection with moral reasoning. However, most of the focus has been on accounting for the so-called "side effect-effect" (Knobe, 2003), in which participants' explicit judgments of intentionality with respect to the effects of others' actions sometimes appear to be influenced by their moral evaluation of these effects. In so far as the side effect-effect is a real phenomenon (there is some debate about this), it seems that our judgments of whether an effect of an agent's action is intended versus merely foreseen may be influenced by moral considerations. However, this does not prevent the causal relationship from going in the other direction as well, and my concern here is with how the intend/foresee distinction may be cognised when it does indeed go in this other direction. Moreover, explicit judgments of intentionality of the sort at issue in the debate over the side effect-effect may potentially come apart from unconscious judgments of intentionality of the sort at issue here, perhaps being the product of quite different cognitive processes.

[5] This is so even in the case of Mikhail's (2011) otherwise extremely detailed and worked out account of how the DDE (and other candidate moral principles) may feature in moral cognition and be applied by our moral reasoning faculty to particular moral problems. Though Mikhail provides complex structural descriptions of the actions that take place in his trolley scenarios, and attempts to formalise how we unconsciously apply the intend/forsee distinction to mental representations of these act-descriptions via the notions of I-generation and K-generation, the precise nature of this distinction is not spelled out. In particular, the categorisation of particular acts and effects in his scenarios as intended or merely foreseen by the relevant agents is just stipulated. No general principles are given for how we go about determining *when* an act or effect is intended versus merely foreseen, only an account of what inferences are made when we do in fact make such determinations.

meant to rule against (e.g., the Footbridge case), we find that it ends up also ruling against actions it is meant to rule in favour of (e.g., the Switch case). The problem for proponents of the DDE hypothesis in moral psychology is that by failing to provide an account of how they think the intend/foresee distinction is drawn by the cognitive processes underlying moral judgment, they have so far failed to show that a tacit psychological version of the DDE can in fact explain the current empirical data on folk moral intuitions. Indeed, it is possible that this data may actually provide evidence *against* the DDE being a fundamental component of human moral psychology.

By the same token, critics of the DDE hypothesis in moral psychology face a similar problem. In particular, Waldmann and colleagues (Waldmann and Dieterich, 2007; Waldmann and Wiegmann, 2010) have presented empirical studies purporting to contradict the DDE hypothesis and support an alternative hypothesis about the cognitive processes underlying folk moral judgment, but have been equally vague in their articulation of the intend/forsee distinction, and consequently have failed to show that the DDE hypothesis *cannot* account for the results of these studies.[6]

I will first highlight the problem of precisely formulating the distinction between intended and foreseen consequences and its implications for interpreting the current empirical data on folk moral intuitions. I will then distinguish between three different approaches to this problem that have been taken by proponents of the DDE in normative ethics: two different kinds of attempt to explicate the intend/foresee distinction so that it is adequate to support the DDE (so-called "closeness" accounts and accounts that employ what come to be known as a "strict" notion of intention), and Warren Quinn's recasting of the DDE in terms of the distinction between "direct" and "indirect agency". I will show that when taken as a claim about our moral psychology, these different accounts entail quite different psychological models of how the folk reason about the intentions of other agents and the causal structure of their actions when assessing the moral status of these actions, and that they lead to different empirical predictions about what people's

[6] One critic of the DDE hypothesis who does seem to be aware of this issue, but does not discuss it in detail, or take it into account when evaluating his own arguments against the hypothesis, is Joshua Greene, who notes that "there is an interesting psychological problem here: namely to understand the mechanism that parses events in these contexts" (2013, p377; see also Greene et al., 2009, p370). Greene's own characterisation of the DDE hypothesis comes closest to that I discuss in Section 3.3, but he does not seem to be aware that this is but one of several different ways of formulating the hypothesis.

moral intuitions should be in particular cases. Based on the current empirical data, I will argue that a psychological version of Quinn's formulation of the DDE seems to be by far the more empirically plausible, and, moreover, that adopting a Quinn-style formulation helps to diffuse some, though not all, of the empirical criticisms of the DDE hypothesis levelled by Waldmann and colleagues. In so doing, I will point towards some potentially fruitful avenues for future empirical work in this area.

2. Can the DDE explain the moral intuitions it is meant to explain?

John Mikhail, the leading defender of the claim that the DDE is a fundamental feature of our moral grammar, formulates the principle as follows:

> [A]n otherwise prohibited action, such as battery or homicide, which has both good and bad effects may be permissible if the prohibited act itself is not directly intended, the good but not the bad effects are directly intended, the good effects outweigh the bad effects, and no morally preferable alternative is available. (Mikhail, 2011, p149)

Mikhail's formulation of the DDE reflects a standard formulation of the principle in normative ethics, which relies on the distinction between *intended* and merely *foreseen* consequences of an action: when the other conditions (such as the proportionality condition) are met, an act, such as one involving battery or homicide, which would otherwise be impermissible, may be permissible, so long as the agent does not intend to commit battery or homicide as a goal of the action ("the prohibited act itself is not directly intended"), and the agent "directly intends" only the good consequences of the action, not the bad ones.

But what exactly does the distinction between "directly intending" and merely "foreseeing" the bad effects of an action amount to? For protagonists in the debate surrounding the DDE in normative ethics, this question amounts to that of when individuals should be seen as intending particular consequences of their actions as opposed to merely foreseeing them. Critics of the DDE (e.g., Davis, 1984; Bennett, 1995) have argued that there is no principled account of intention that can do the work that proponents of the DDE need it to do.

To illustrate the problem, consider the standard Footbridge case, where the agent

pushes the victim (the large man) off the footbridge in front of the trolley in order to save five people. On standard DDE accounts of this case, it is meant to be clear that the agent "directly intends" the bad consequences of this action (the large man being harmed), as well as the good consequences (the five people on the main track being saved), and this is supposed to contrast with the Switch case, where it is meant to be clear that the agent does not directly intend the bad consequences of the action (harming the person on the side track). But consider the following, intuitively plausible, account of which consequences of an agent's actions are intended versus merely foreseen by the agent (adapted from Bennett, 1995):

> (1) An agent S intends all those consequences of an action A that S believes will occur as a result of doing A and which explain why S does A, while all the other consequences of A that S believes will occur are merely foreseen.[7]

On this account of the intend/foresee distinction, the agent in the Footbridge case would be seen as intending that the five people be saved, intending that the large man be put in front of the trolley, and intending that this stop the trolley. This is because all of these consequences of the agent's pushing of the large man must surely figure in an explanation of why the agent pushed him. However, the agent would *not* be seen as intending to cause any harm to the large man. Though the agent must surely have believed that these harmful consequences would result from this action, they do not figure in an explanation of *why* the agent pushed the man, at least if we assume (as DDE accounts of this case normally do) that the agent only pushed him in order to save the five, so the agent would be quite happy if the large man was miraculously unhurt by the interaction with the trolley. Hence, if the above account of the intend/foresee distinction is correct, the DDE would not deliver the negative verdict it is meant to deliver in this case. Instead, it would suggest it was just as permissible for the agent in the Footbridge case to push the large man in front of the trolley as for the agent in the Switch case to throw the switch. In both cases, the harm caused in service of a good end was merely foreseen, not intended.

[7] A similar account of the intend/foresee distinction to (1) can be found in Bratman's (1999) influential account of intention, according to which an agent intends all those consequences of an action A that the agent is "committed" to bringing about by doing A.

Explications of the DDE sometimes explicitly state that if an agent intends a particular end, the agent must also intend the known means. However, this is clearly not an adequate response to the problem, since the issue is partly that of determining *what* effects of an agent's action should be considered part of the agent's means. What critics of the DDE claim is that while it seems clear that the large man being put in front of the trolley in such a way as to stop it is part of the agent's intended means to save the five in the Footbridge case, it is not all clear that the *harm* caused to the large man is part of the means and not just a foreseen side effect of his being put in front of the trolley.

In response to these types of criticisms, proponents of the DDE have tended (though, as we will see, not universally) to reject accounts of the intend/foresee distinction such as that found in (1), and have tried to formulate an alternative account of intention that ensures that the DDE can actually deliver the verdicts it has been thought to deliver. We will look at some of the specific proposals that have been made for explicating a relevant notion of intention in the next section. We'll also look at a quite different way of formulating the DDE, proposed by Warren Quinn, which does not rely on the distinction between intended and foreseen consequences. But, first, it is important to emphasise that the nature of the challenge in clarifying the intend/foresee distinction is importantly different for proponents of the DDE hypothesis in moral psychology. For the principle's normative proponents, the challenge is to give a principled account of when agents *should* be seen as intending something versus merely foreseeing it, and then to show that the DDE does deliver the right verdicts in particular cases. Thus, of primary importance here is the issue of what is the *correct* account of intention, and hence of what individuals may correctly to be said to intend in various circumstances. In contrast, for the principle's psychological proponents, the challenge is to provide a plausible account of how the cognitive processes underlying folk moral judgment attribute intentions to other agents when applying the principle—irrespective of whether these particular attributions are correct or not—and to show empirically that patterns of moral judgment conform to what we would expect to see if a distinction between intended and merely foreseen consequences plays an important role in moral judgment. Thus, for the moral psychologist, the question of how the intend/forsee distinction should be formulated is a purely *descriptive* question about the cognitive processes underlying folk moral

judgment. Thus, no particular account of what intentions *actually are* needs to be defended.

Most probably, (1) is not a plausible account of how the intentional character of the actions of other agents is conceived by the cognitive processes underlying folk moral judgment. Nonetheless, it should be clear from the above discussion that proponents of the DDE hypothesis in moral psychology should be as concerned to clarify the distinction at the heart of their version of the principle as the DDE's normative proponents, since these problems bear directly on how we should interpret the current stock of experimental results: if something like (1) were correct as an account of moral psychology, then the DDE hypothesis would obtain *no* empirical support at all from the current set of trolley studies. Indeed, other things being equal, the fact that participants generally regard the action in the Footbridge case as impermissible would seem to constitute evidence *against* the DDE being part of our moral psychology.

3. Three responses to the problem in normative ethics

Let us now consider three quite different approaches to the problem surrounding the intend/forsee that have featured in the literature in normative ethics. Once we have done this, we can then think about which approach, when translated into a claim about folk moral psychology, may be better supported by the current empirical data, and how future empirical work could test for the presence or absence of the DDE in our psychology.[8]

As we have seen, standard formulations of the DDE in normative ethics draw a moral distinction between causing harm intentionally and merely foreseeing that harm will result as a consequence of one's action. In response to the concerns described in the previous section, proponents of the DDE have typically tried to offer an account of intention that blocks the possibility of viewing agents in cases like the Footbridge case as merely foreseeing the harm that results from their actions. Several different types of approach have been adopted, but I want to focus on two families of accounts: what we

[8] Of course, something that needs to be kept in mind, particularly in what follows, is given that the DDE is not a comprehensive moral doctrine—and thus could only be *one* component of a folk morality—judgments that deviate from the DDE might just indicate the influence of other moral principles. We should be mindful of this both when trying to distinguish between different versions of the DDE hypothesis, and distinguishing between some version of the DDE hypothesis and rival hypotheses of sort considered later in the paper.

can call "closeness" accounts, and accounts that employ what has become known as a "strict" notion of intention. In Section 3.3, we will look at Quinn's attempt to recast the DDE in quite different terms.

3.1 Closeness accounts

As illustrated by (1), most traditional philosophical accounts of intention regard the question of what intentions an agent has in performing a given action as at root a question about what beliefs and desires the agent has and the causal or explanatory role that those beliefs and desires play in the generation of the action. According to closeness accounts, however, a distinction is to be drawn between what we might call the *cognitive intentions* of an agent, which is captured by this traditional view (and perhaps by (1)), and a *thick* notion of intention that relates to what an agent should be seen as intending given the nature of the action, whether or not the agent actually has any combination of beliefs and desires that correspond to these intentions. It is this thick notion of intention that is claimed to be morally relevant for the purposes of applying the DDE. The idea is that, in certain circumstances—such as the Footbridge case—an agent may be properly said to intend to cause harm to another person, even if the agent has no combination of beliefs and desires that would, on the traditional cognitive view, be regarded as constituting an intention to cause the harm, and, on reflection, the agent fails to view the harm as intended. Following a suggestion from Phillipa Foot (1967), the idea is that actions such as pushing the large man in front of the trolley are just "too close" to the harm that results from these actions for this harm to be plausibly seen as merely a foreseen side effect. One cannot plausibly say that the agent intends that the man be put in front of the trolley, and intends that this stop the trolley, without also saying that the agent intends the harm that results from this action.[9]

The challenge for closeness accounts is, first, to justify employing a thick notion of intention in moral evaluation that may come apart from what is going inside the agent's head, and, second, to specify exactly how and when particular harms are "too close" to

[9] Arguably, tying the DDE to more objective features of the causal structure of the agent's action, rather than to features of the agent's psychology, is more consonant with how the doctrine has been understood historically, particularly in the Catholic intellectual tradition (for discussion of the DDE in Catholic moral theology, see Connell and Kaczor, 2013).

particular actions for their being brought about to be plausibly seen as merely foreseen and not intended, in a way that distinguishes between the right kinds of cases.

One attempt to solve the second challenge employs the notion of *act identity* (Anscombe, 1963; Davidson, 1980). To illustrate this account, consider the much-discussed Craniotomy case. In this case, a doctor intends to save the life of a pregnant woman, whose foetus is trapped in the birth canal. The only way to do this is by performing a craniotomy on the woman's unborn foetus: crushing the foetus' skull and thereby killing it, but also allowing it to pass through the birth canal without killing the mother. This kind of action has traditionally been thought to obtain no justification from the DDE. This because it is meant to be clear that the foetus is harmed intentionally as a means to saving the mother.[10] However, the worry raised by critics is that the same reasoning used by proponents of the DDE to justify actions such as that in the Switch case can be used here too, since it could be argued that the doctor merely foresees that the foetus will be killed by the craniotomy and does not intend this harmful effect as a means to the goal of saving the mother. On any charitable interpretation of the doctor's state of mind, the key motivating desire is save the mother, not to harm the foetus, and it is the belief that performing the craniotomy will save the mother that explains why the doctor performs the procedure, not the belief that the craniotomy will harm the foetus (of course, the action may still fail other conditions of the DDE, such as the proportionality condition).

Proponents of the act identity account claim that the same action may fall under various descriptions. In the Craniotomy case, one description of the action is crushing the skull of the foetus. However, given various facts about the world (such as the nature of skulls and what happens to foetuses when their skulls are crushed), another equally appropriate description of this act is *killing* the foetus. After all, it is not as if some *other* action must be performed after the act of skull crushing in order for the foetus to be killed—it just is the act of skull crushing that kills the foetus. Thus, the claim is that it would be bizarre to say that an agent can intend to perform one of these actions without also intending to perform the other, since they are just different descriptions of the same

[10] This does not necessarily mean that the action must therefore be impermissible, only that if it is permissible for the doctor to perform the craniotomy, it is not the DDE that explains why.

act. The same is meant to hold for the Footbridge case: given various facts about what happens to people put in front of moving trolleys, another equally appropriate description of the act of pushing the large man in front of the trolley is the act of killing him. Thus, if an agent intended to push him in front of the trolley, it would be bizarre to say that the agent did not also intend to harm him on this account.

The problem with this act identity account, however, is that it seems to lead the DDE to rule against many other cases that it has been thought to justify (Davis, 1984). Consider the Switch case. Here it seems just as plausible as in the Footbridge and the Craniotomy cases to say that another equally appropriate description of the act of flipping the switch to divert the trolley onto the side track is the act of killing the man on the side track. Again, no other action has to be performed after this one in order for the man to be killed. Given the particular facts of the case, it is this act that guarantees that the man will be killed. Thus, plausibly, the act identity account of the intend/forsee distinction would lead the DDE to rule against the agent's actions in this case as well.

Another, more recent, and more plausible, attempt to define a notion of closeness comes from William Fitzpatrick (2006). His account trades on the relationship between *states of affairs*, rather than acts. The idea is that given what agents know about the world, the relationship between certain states of affairs is not merely a causal one—i.e., a case of the one causing the other—but rather a *constitutive* one: the bringing about of the one just *is* the bringing about of the other. As Fitzpatrick puts it:

> [I]f the relation between two states of affairs is known to the agent, natural and constitutive rather than merely causal, then we cannot properly speak of an agent's intending the one while merely foreseeing but not intending the other. (W. Fitzpatrick, 2006, p603)

On this account, the man being harmed must be said to be intended and not merely foreseen by the agent in the Footbridge case, since, given what we know about moving trolleys and what they tend to do to people who are put in front them, the relationship between the state of affairs of the large man being put in front of a moving trolley in such a way as to stop it and the state of affairs of him being hurt is not merely a causal relationship but rather a *constitutive* one: given this contingent state of the world, the man's being put in front of the trolley doesn't merely *cause* his being hurt, it *constitutes*

his being hurt. Thus, assuming that the agent knows what the world is like, if the agent intends to put the man in front of the trolley, the agent must also be said to intend to harm him. The same is meant to be true for the Craniotomy case: the state of affairs of the foetus' skull being crushed is constitutive of its being killed. In contrast, in the standard Switch case, Fitzpatrick's claim is that the relationship between the state of affairs of the agent flipping the switch to move the trolley onto the side track and the state of affairs of the man being hurt and killed is merely a *causal* relationship, *not* a constitutive one. Though the agent knows that the flipping of the switch will bring about this harmful state of affairs, in no way is the one state of affairs constitutive of the other. Fitzpatrick allows, on his account, that the act of flipping the switch is identical to the act of killing the man. Nonetheless, the agent can still properly be seen as intentionally bringing about the one state of affairs, while merely foreseeing the other as an unintended side effect.

For current purposes, I am not interested in whether Fitzpatrick's account (or any other kind of closeness account) provides a satisfactory articulation of the intend/foresee distinction at the heart of the DDE as a normative principle. Rather, what I want to take from this sort of account is a pointer towards one way of thinking about how we might go about attributing intentions to agents in these sorts of cases. When converted to a claim about our moral psychology, what this sort of account suggests is that in applying the DDE we mentally analyse the causal connection between the consequences of an action and the means that the agent uses to bring about the goal of the action. For instance, when an agent is mentally represented as intending that a particular state of affairs be brought about and this state of affairs is mentally represented as being constitutive of another state of affairs that is the harming of a victim, the agent is mentally represented as intending the harming of the victim. In contrast, when this harmful state of affairs is mentally represented as merely a causal consequence of (and not constitutive of) the state of affairs that the agent is represented as intending to bring about, the agents is represented as merely foreseeing the harm. Naturally, this would require some (complex) criteria for distinguishing between constitutive and causal links between states of affairs.

3.2 Strict intention accounts

In contrast to closeness accounts, the second family of intend/foresee accounts of the

DDE does not rely on a thick notion of intention. As Alison Hills has put it, the concern for many proponents of the DDE about such a thick notion of intention is that "[i]ntentions are intensional: however ''close'' X is to Y, even if X is identical with Y, it is possible for an agent to intend that X and not to intend that Y" (2007, p265). Hills gives the example of a strategic bomber who intends to destroy the ball-bearing factory, but does not intend to destroy the largest factory in the city, even though these factories are one and the same. Thus, instead of utilising some notion of closeness, the goal for these theorists has been to articulate a notion of intention that is adequate to explicate the DDE, but which does not depart from the traditional idea that to talk of an agent's intentions is to talk about some combination of the agent's actual beliefs and desires.

A recent example of such a "strict" account of intention comes from Lawrence Masek:[11]

> I contend that an effect is intended (or part of the agent's plan) if and only if the agent A has the effect as an end or believes that it is a state of affairs in the causal sequence that will result in A's end. Any other effect is unintended, even if A foresees it with certainty. (Masek, 2010, p569)

Note the specific reference to the agent's actual beliefs and desires: for the agent to intend some effect of an action, this effect must either be something that the agent desires to bring about as an end, or be something that the agent believes is part of the causal chain initiated to bring about their end. Masek's notion of a causal chain is left rather vague, but he does claim that the harm to the large man is part of the causal chain that will result in the agent's end in the Footbridge case. In contrast, the harm to the man on the side track is not part of the causal chain that will bring about the agent's end in the Switch case, but is rather part of a different causal chain that was initiated, but nonetheless branches off from the one that brings about the agent's end. Since the agent in the Footbridge case must believe that the harm to the large man is on the causal path to achieving the end of saving the five, it is indeed, on Masek's account, something that the agent intends to bring about in initiating the action, whereas the harm to the man on the sidetrack is not intended and merely foreseen by the agent in the Switch case. Thus,

[11] As Masek points out, the term "strict" was originally used as a term of abuse by proponents of closeness accounts to refer to accounts of intention that they judged to be too narrow, but it has recently been taken on by proponents of a family of accounts of the DDE that depart from closeness accounts.

unlike (1), Masek's account of how to determine whether an effect is intended versus merely foreseen allows us to anchor an agent's intentions to some combination of actual beliefs and desires, while still allowing (he argues) the DDE to deliver the moral verdicts that it has been thought to deliver.

Again, for the purposes of this paper, I am not interested in the normative adequacy of such a strict intention account of the DDE. Rather, what I want to take from it is a pointer towards a contrasting view of how the intend/foresee might be psychologically implemented to that suggested by closeness accounts. What Masek's account suggests is that when determining whether a particular effect of an agent's action is intended versus merely foreseen, the focus is on what the agent is mentally represented as having as an end, and what the agent is held to believe about the states of affairs that are on the causal pathway to bringing about the agent's end. As with closeness accounts, the causal structure of the action plays a vital role, but different aspects of this causal structure are what do the work in determining whether particular effects of an agent's act are intended versus merely foreseen.

3.3 Quinn's account

In contrast to the two families of accounts just discussed, Warren Quinn's (1993) account of the DDE seeks to distance the principle from traditional formulations in terms of the distinction between intended and foreseen consequences. The reason for this is that Quinn broadly accepts something like (1) as an account of the intend/foresee distinction, and is concerned that applications of the DDE do not ascribe intentions to agents for which there are no corresponding mental states in the mind of the agent. Instead, Quinn argues that the DDE should be based around a distinction between actions where there is an intention on the part of the agent to *involve* a person in some event that causes them harm as a means or an end—what Quinn calls "directly harmful agency"—and actions that result in harm, but without such intended involvement of the persons concerned as a means or an end—what Quinn calls "indirectly harmful agency". The idea is that the DDE rules against the former type of agency, which causes harm by using people as a means, but may permit the latter type, which causes harm merely as a side effect. Quinn's version of the DDE thus has an explicitly Kantian flavour to it: it rules against

deliberately *using* people against their will as a means to achieving some goal in a way that the agent knows will cause them harm.

Importantly, for Quinn, for the agency to be directly harmful, it is does not matter whether or not the agent intends to cause harm to the person. Rather, what the agent must intend is to *involve* them in some event as a means to achieving some goal that the agent knows will result in harm to the person. This allows for a potential disconnect between the agent's intentions and the resultant harm, which is precisely what both closeness and strict intention accounts want to get rid of. To see this, consider Jonathan Bennett's (1995) example of a terror bomber who claims not to actually intend to cause harm to the civilians killed during a bombing raid on a civilian area aimed at destroying enemy morale: the bombing of them was not motivated by an intention to kill them, but rather by an intention to make them *appear* to be dead in order to destroy enemy morale—thus, the bomber would be quite happy if the civilians survived, so long as they appeared to be dead long enough for their "seeming deaths" to have the right effect on enemy morale. Closeness accounts of the DDE try to block this move: maybe the terror bomber was not actually motivated by any combination of mental states corresponding to a cognitive intention to cause harm to the civilians, but because of the closeness of the bomber's intended means (bombing them) to this harm, the bomber nonetheless *did* intend to cause them harm. Quinn, on the other hand, allows that in this case the terror bomber did not in fact intend to cause harm to the civilians. However, what the bomber surely did intend was to *involve* them in an event (their being bombed) that he or she knew was going to (or was at least highly likely to) result in harm to them. Thus, on Quinn's account, this is a case of directly harmful agency, and this is what the DDE, on his formulation, rules against.[12]

The essence of Quinn's account is a counterfactual principle: is the particular involvement of the person or persons who suffer harm as a result of the relevant action causally *necessary*, given the constraints of the particular situation, for the achievement

[12] Quinn's discussion of the DDE suggests that the principle specifies when actions are to be regarded as impermissible. However, I agree with Mikhail (2011, p152) that the principle is better understood not as specifying the conditions under which actions are impermissible, but rather conditions under which otherwise impermissible actions may be permissible. Though this may have implications for considering the normative adequacy of Quinn's formulation (in particular, it would pose difficulties for his Kantian justification for the principle), it does not, I think, have any bearing on my use of his formulation here.

of the agent's goal? If this involvement is necessary, then the case is an example of directly harmful agency. In the Footbridge case, the involvement of the large man in the event that causes him harm—the stopping of the trolley—is clearly necessary, for without the large man being involved the trolley would not be stopped and the five people would not be saved. The same is true for the Craniotomy case, for, again, without the involvement of the foetus the goal of saving the mother would not be achieved: given the constraints of the imagined situation, the only way to save the mother is to crush the foetus' skull, and that is to deliberately involve it in an event that causes it harm. However, this is not so in the Switch case, because there is no such necessary involvement of the person on the side track. If the person on the side track had not been there, or was able to get off the side track in time, the agent's means would have been exactly the same, and the agent's goal would still have been realised.

Again, the normative adequacy of Quinn's version of the DDE is not what is at issue here (see Fischer et al., 1993; W. Fitzpatrick, 2006 for discussion). What a psychological version of Quinn's DDE suggests is that the focus in applying the principle is on whether or not an agent is judged to have intentionally *involved* a person in some event that the agent knows will result in harmful consequences as a means to achieve the ultimate goal of the action. The notion of intentional involvement here is counterfactual: an agent is mentally represented as intending to involve a person in an event as a means to an intended end if, given the constraints of the situation, the involvement of that person is mentally represented as causally necessary for attaining the intended end. If this involvement is not mentally represented as causally necessary for attaining the intended end, then the agent is not mentally represented as intentionally involving the person as a means. In contrast to closeness and strict intention accounts, there need be no mental representation of what the agent intends or merely foresees with respect to the harm caused to the victim.

4. Empirically distinguishing between the three models

We can now see that when we compare closeness accounts, such as Fitzpatrick's, strict intention accounts, such as Masek's, and Quinn's account of the DDE, and consider how they might be translated into claims about our moral psychology, we get different

psychological models of which specific elements of an action and the agent's state of mind may be taken into account when the DDE is applied in particular cases. These differences are subtle, but important, and they lead to different empirical predictions about what our moral intuitions should be in particular cases if the DDE is indeed a fundamental component of human moral psychology.

To illustrate the different empirical implications of these models, consider the following trolley scenario (modified from Waldmann and Dieterich, 2007): a runaway trolley is about to hit a bus with five people trapped inside that has become stuck on the tracks. The agent can flick a switch to divert the trolley onto a side track, which loops back to the main track before the bus. There is a car on the side track before it loops back with a single person trapped inside, who will be killed by the trolley if it is diverted. However, the car will stop the trolley, and the five people will therefore be saved. Now let us consider two versions of this scenario:

I. The agent knows that there is a single person inside the car.
II. The agent knows that there is a single person inside the car, and knows that the person being inside is *necessary* for stopping the trolley: the person's weight is required in order for the car to stop the trolley; if the person were not there, the trolley would continue on and kill the five people in the bus.

Is it morally permissible for the agent to flick the switch to divert the trolley onto the side track? It would that seem that according to an excessive closeness account of the DDE along the lines of Fitzpatrick's account, both of these actions are equally impermissible. On this formulation, the focus is on the causal and constitutive links between the states of affairs brought about by the action. In both I and II, the relationship between the state of affairs of the car being used to stop the trolley and the state of affairs of the person inside the car being hurt is plausibly a constitutive one. Thus, given that the agent clearly intends to use the car in this way as a means to save the five people in the bus, the agent must also be said to intend to cause harm to the person in the car as a means to save the people in the bus. Hence, if this version of the intend/foresee distinction at the heart of

the DDE is ingrained in our folk moral psychology, we should see no significant difference in permissibility ratings from experimental participants for I and II.

A strict intention formulation, along the lines of that suggested by Masek's account, would seem to produce a similar prediction, since the harm caused to the person in the car is presumably on the causal pathway to the attainment of the good end in both I and II. It would seem that in both cases the causal pathway to saving the five goes like this: agent flicks the switch, the trolley is diverted onto the side track, the car on the side track is struck by the trolley, killing the person inside, but also stopping the trolley and saving the five.

However, if a Quinn-style formulation of the DDE is part of our moral psychology, I may be judged as permissible, and II as impermissible. On this formulation, what is important is the counterfactual test for intentional involvement: given the constraints of the situation, is the involvement of the person in the car causally necessary for the achievement of the goal of the action, and does the agent know this? In II, the involvement of the person in the event that causes them harm is explicitly stated to be necessary, for the absence of the person would subvert the achievement of the agent's goal, and this is known by the agent. Hence, the action in II would be a case of directly harmful agency. However, this is not the case in I, so this action could be seen as a case of indirectly harmful agency, and so may be permissible.

For another example of how these accounts may come apart empirically, consider Mikhail's Man-In-Front case (2011, p108). A runaway train is about to hit five men on the main track, but the agent can flick a switch to divert the train onto a side track that loops back to the main track before the five men. On the side track is a heavy object that will slow the train enough for the men on the main track to escape. However, there is a man on the side track in front of the heavy object with his back turned. If the agent flicks the switch, the five men will be saved, but the man on the side track will be killed. A Fitzpatrick-style formulation would seem to predict that participants should judge this action to be permissible (which is what the majority of Mikhail's participants did in fact do). This is because the state of affairs of the large object slowing the train is not constitutive of the man on the side track being hurt and killed. Thus, the agent would be mentally represented as intending to bring about this state of affairs as a means to saving

the five, but only as foreseeing the harm caused to the man on the side track. A Quinn-style formulation of the DDE would seem to produce the same prediction. The involvement of the man on the side track is clearly not causally necessary for the achievement of the agent's goal. Thus, the agent would not be mentally represented as intending to involve the man in an event that causes him harm. In contrast, a Masek-style formulation would seem to predict that participants should judge the action as impermissible. The causal sequence appears to run like this: agent flicks the switch, the train is diverted onto the side track, the train hits the man on the side track and kills him, the train hits the large object, slowing it down, the five are saved. Thus, since the harm caused to the man is clearly on the causal pathway that leads to the attainment of the agent's end, the agent should be mentally represented as intending to harm the man on the side track.

Clearly, much more needs to be said to fully clarify these three different psychological models of how the distinction at the heart of the DDE may be cognised. But from what little I have said here we can already see that applying these different accounts of the DDE as a normative principle to the problem of formulating a psychological version of the DDE may generate quite different empirical predictions about participants' responses to particular cases. This suggests fruitful lines of research to determine which, if any, of these formulations of the DDE hypothesis is empirically adequate. What I want to do now is suggest some preliminary reasons for regarding a Quinn-style formulation, which uses a counterfactual test for determining intentional involvement, as the most empirically adequate based on the results of the studies that are already published. In so doing, I will diffuse some of the recent empirical criticisms of the DDE hypothesis, and highlight a particularly promising avenue for future research.

5. Some preliminary reasons to favour a Quinn-style formulation

Contra the claims of Mikhail and others, Waldmann and colleagues (Waldmann and Dieterich, 2007; Waldmann and Wiegmann, 2010) have claimed that it may not be the DDE that is really doing the work in producing contrasting folk moral intuitions in response to hypothetical moral dilemmas, such as the Footbridge and Switch cases, but rather a cognitive phenomenon that they call "intervention myopia". The act described in

the Switch case involves an intervention on the thing that causes the harm—what Waldmann and Wiegmann (2010) call the "threat"—which, in this case, is the trolley. However, the act described in the Footbridge case involves an intervention on a "victim" of harm—the large man—who suffers harm as a result of the intervention. According to Waldmann and colleagues, this difference in the locus of intervention focuses our attention on different aspects of the two cases: in the Switch case, our attention is focused on the harm that will befall the five men on the main track if the threat (the trolley) is not diverted. In such threat intervention cases, we are, they argue, more inclined to discount the harm that the intervention will cause others (the man on the side track).[13] However, in victim intervention cases, such as the Footbridge case, our attention is focused on the harm that is going to be caused to the victim (the large man) by the intervention, so we are less likely to discount this harm against the harm that will be caused to others by inaction. We therefore have a harder time permitting the action performed in victim intervention cases than in threat intervention cases.[14]

[13] They do not claim that we discount it altogether, but rather that it recedes into the background relative to the harm that will be caused by not acting.

[14] Greene advances a similar hypothesis to that of Waldmann and colleagues. According to his "modular myopia" hypothesis:

> We have an automatic system that "inspects" action plans and sounds the alarm [i.e., delivers negative moral evaluations, such as impermissibility judgments] whenever it detects a harmful event in an action plan (e.g., running someone over with a trolley) But […] this action-plan inspector is a relatively simple, "single-channel" system that *doesn't keep track of multiple causal chains*… Instead, when it's presented with an action plan for inspection, it *only sees what's on the primary causal chain*. (Greene, 2013, p234 [italics in original]).

Hence, according to Greene, the reason that we intuitively distinguish between the moral acceptability of the "action plans" in the Switch and Footbridge cases, is that this automatic system can't "see" the harm caused to the man on the side track in the former case, since it takes place on a causal chain that branches off from the primary one leading to the saving of the men on the main track, but *does* "see" the harm caused to the large man in the latter case, since it takes place on the primary causal chain. Like Waldmann and colleagues, Greene argues that the nature of this automatic system leads us to make judgments consistent with the DDE in cases like these—since it cannot "see" harmful side effects—but, in others, our judgments come apart from the DDE. Greene argues that we also have another system that engages in more reflective moral reasoning, which may sometimes conflict with, and occasionally override, the judgments issued by the automatic system. To be clear, Waldmann and colleagues and Greene need not be seen as completely downplaying the role of the agent's intentions in (automatic) moral judgments. For instance, Greene argues that the automatic system can take intentions into account, such that intentional harms occurring on the primary causal chain are judged to be worse that unintentional ones. Similarly, (as Greene et al., 2009, p369 point out) Waldmann and colleagues' approach seems to imply that agents are seen as intending to act on victims versus threats, and allows for a possible moral distinction between intentional and accidental harms.

In order to distinguish this intervention myopia hypothesis from the DDE hypothesis, Waldmann and Dieterich (2007) constructed a case in which a runaway trolley is about to hit a bus with ten people trapped inside. The only way to stop this is by flicking a switch to turn the trolley onto a side track where there is another bus with two people trapped inside, who will be killed by the trolley. This case is meant to be analogous to the standard Switch case. They compared participants' responses to this case with a case in which the side track loops back to the main track, so if the trolley does not hit the bus with the two people inside, it will continue on to hit the bus with ten people inside. Waldmann and Dieterich found no significant difference between the ratings of the actions in these two cases: subjects generally approving of both actions, even though the DDE hypothesis, according to Waldmann and colleagues, predicts that the action in the second (Loop) case would be judged to be less acceptable than the action in their version of the Switch case. In the Loop case, they claim, the harm caused to the people on the side track is an intended means to the end of saving the ten people on the main track. Since both of these cases involve threat rather than victim intervention, Waldmann and colleagues consequently argue that this supports the intervention myopia hypothesis—it is more likely to be the distinction between threat and victim intervention that is doing the work in explaining the different verdicts on the Footbridge and Switch cases than the DDE distinction between intended and merely foreseen consequences.[15]

Here, however, issues about the precise formulation of the distinction at the heart of the DDE arise: it is actually far from clear *what* the DDE rules in Waldmann and Dieterich's Loop case. Consider Quinn's account of the DDE. Since the people on the side track are in a bus, and it is plausibly the bus that does the causal work in stopping the trolley in the Loop case, it would seem that in this case and their version of the Switch case the involvement of the two people on the side track is *not* causally necessary for achievement of the goal of saving the ten—at least on the assumption that an empty bus would be just as good for stopping the trolley. Thus, both cases would be instances of

[15] One potential methodological problem here is that Waldmann and colleagues asked participants about whether the agent "should" or "should not *act"* (my emphasis) in the way described in the relevant situation, which is rather different from asking whether the action is "permissible" or "impermissible" (the question asked in most of the experiments that have been claimed to support the DDE hypothesis, such as those of Mikhail). For instance, one can imagine cases in which one might judge that a person's action is permissible, but also judge that the person should not perform that act. Whether or not this is a significant confound is unclear.

indirectly harmful agency. Waldmann and Dieterich's results therefore appear to be perfectly consistent with a Quinn-style formulation of the DDE hypothesis, centred on the counterfactual test for intentional involvement.

These results do, however, seem to tell against intend/foresee formulations of the DDE hypothesis based either on Fitzpatrick's closeness account or Masek's strict intention account. In the Loop case, the state of affairs of the bus with two people inside being hit by the trolley is presumably (given the constraints of the situation) constitutive of the two people being hurt and killed. Hence, the agent should be mentally represented as intending to bring about the latter state of affairs as much as the former, rendering the action of diverting the trolley impermissible on a Fitzpatrick-style formulation of the DDE. The harm caused to the two in the bus is presumably also on the causal path to the attainment of the agent's end in this case, implying that the action should also be judged to be impermissible on a Masek-style formulation.

Waldmann and Dieterich's results therefore seem to provide some empirical motivation for adopting a Quinn-style formulation of the DDE hypothesis over alternative formulations based on a Fitzpatrick-style closeness account or a Masek-style strict intention account.

The results of another experiment ran by Waldmann and Wiegmann (2010) might seem to tell against the DDE hypothesis, if a Quinn-style formulation is adopted. In this experiment, participants were told that a runaway train is about to hit another train on the main track with five workers trapped inside, who will be killed. There is a parallel side track, which connects to the main track via a connecting track. In all conditions a single passenger is inside the threatening train, but in a compartment at the back of the train. If the agents in the control centre do nothing, the threatening train will cause the deaths of the five workers, but the single passenger at the back of the threatening train will survive unhurt. In Condition I, another train can be redirected from the side track to hit and derail the threatening train. This would save the five workers, but kill the passenger on the threatening train. In Condition II, the single passenger in the threatening train is standing near the brake system of the train, but has no idea how to use it and cannot be communicated with. The agents in the control centre can, however, redirect the train on the side track to hit the threatening train, which will cause the passenger to be knocked

against the lever controlling the brakes of the train. The train will therefore be stopped and the five workers saved, but the passenger will be killed. Waldmann and Wiegmann report participants giving high ratings to the actions in both conditions, and argue that this poses serious problems for the DDE hypothesis, since (they claim) the single victim was used as a means to save the five in both cases.

It seems clear, however, that a Quinn-style formulation of the DDE hypothesis is perfectly consistent with the result for Condition I: the involvement of the single victim is not causally necessary for the attainment of the agents' goal in this case.[16] Condition II does seem to pose more of a challenge to a Quinn-style formulation (but also to the other two formulations discussed here), since the involvement of the person in the event that causes them harm is explicitly stated as being causally necessary.[17]

This serves to highlight an interesting feature of the current empirical literature, and with it a promising avenue for future research. Consider Mikhail's (2011, p107-108) Loop Track case, in which a runaway train is about to hit five men. The agent can redirect the train onto a side track that loops back to the main track before the five men. However, there is a man on the side track. If the train is diverted, the man will be killed, but the train will be slowed enough for the men on the main track to escape. This action clearly involves an intervention on the threat, but Mikhail's participants were significantly less likely to judge it as permissible compared to the action in the Man-In-Front case (described in Section 4). When put alongside Waldmann and Dieterich's Loop case, this result seems to tell against the intervention myopia hypothesis and for a Quinn-style DDE hypothesis, since it gets directly at the distinction between causally necessary and unnecessary involvement of the victim. However, while Hauser et al. (2007) report a

[16] Waldmann and Wiegmann (2010, p5) do actually seem to acknowledge this point about the victim being causally unnecessary, but don't take it as seriously as they should when it comes to thinking about how the DDE hypothesis ought to be understood.

[17] Another case that seems to pose problems for all three formulations of the DDE hypothesis discussed here is Greene's Collision Alarm case (unpublished research, reported in Greene, 2013, p221-2). Two trolleys are heading down two separate tracks. If nothing is done, the first trolley will kill five workers on the track. However, the second trolley can be redirected onto a side track, where there is one man standing next to an alarm system. If this is done, the trolley will hit and kill the man, but the alarm will be activated, cutting power to the whole network of trolleys, including the first trolley, thus saving the five men. Greene reports a strong majority of participants "approving" of the redirection of the trolley in this case. This would appear to be problematic for a Quinn-style formulation of the DDE hypothesis, since (as Greene describes the case) collision with the man is required to set off the alarm, his involvement is therefore causally necessary for the achievement of the goal

similar result, other researchers have not. For instance, Greene et al. (2009) found no significant difference in responses to versions of Mikhail's Loop Track and Man-In-Front cases.

Greene et al. (2009, p369) attribute the difference in their findings to what they regard as a confound in the original wording of Mikhail and Hauser et al.'s scenarios, where the large man is referred to as a "heavy object" capable of stopping the train in the Loop Track case, but not in the Man-In-Front case (Waldmann and Dieterich provide the same explanation to account for the discrepancy between Mikahil and Hauser et al.'s results and the results from their Loop case). In response, it could be argued that, on a Quinn-style formulation of the DDE hypothesis, the difference in wording simply makes the key moral distinction more apparent: that is to say, the difference in the intended causal role played by the victim. This raises the question of what might happen if the wording of Waldmann and Wiegmann's Condition II were altered to make the causal necessity of the role played by the victim for achievement of the agent's goal more salient to participants than it might otherwise be (similarly for Greene's Collision Alarm case). Interestingly, however, Liao et al. (2012) studied a scenario structurally identical to Mikhail's Loop Track case, in which the causal necessity of the bystander was made highly salient. It included the wording, "if it were not the case that the trolley would hit the innocent bystander and grind to a halt, the trolley would go around [the loop] and kill the five people" (2012, p665). Though the major conclusion of their study was that responses to loop-style cases are influenced by order effects—in particular, participants were significantly more likely to judge the action in their loop case as permissible if presented after cases like the Switch case than the Footbridge case[18]—their participants were generally inclined to judge the action permissible, including when the case was presented first, which does seem to tell against a Quinn-style formulation.

Given these conflicting results, the existing empirical literature therefore seems to be equivocal with respect to deciding between a Quinn-style formulation of the DDE hypothesis and Waldmann and colleagues' intervention myopia hypothesis. Nonetheless, a Quinn-style formulation does seem to be currently the most empirically adequate of the

[18] For interesting discussion of why there may be such order and transfer effects, see Wiegmann et al. (2012) and Wiegmann and Waldmann (2014).

three formulations of the DDE hypothesis considered here. Moreover, adopting such a formulation as a working hypothesis opens up at least one promising area for future research: looking closer at the extent to which the causal necessity of the involvement of the victim in the achievement of the agent's end does indeed play a role in moral judgment, and, in particular, manipulating the saliency of this in the wording of the scenarios.

6. Concluding remarks

I have distinguished between three different ways of formulating the DDE hypothesis in moral psychology, inspired by different accounts of the DDE in normative ethics, and shown that the results of Waldmann and colleagues' experiments, combined with those of Mikhail and others, provide some preliminary empirical motivation for adopting a Quinn-style formulation of the DDE hypothesis over alternative formulations based on closeness accounts, such as Fitzpatrick's, and strict intention accounts, such as Masek's. As I said earlier, much more needs to be done to fully spell out the different psychological models suggested by these different normative accounts of the DDE. In addition, I have focused on only one closeness account and one strict intention account. Other variants might be considered that could potentially inspire more empirically tenable formulations of the DDE hypothesis. Moreover, while I have shown that a Quinn-style formulation takes much of the sting out of the empirical challenge to the DDE hypothesis posed by Waldmann and colleagues, much more research also needs to be done to decide between the most empirically adequate formulation of the DDE hypotheses (whatever that turns out to be) and alternative hypotheses of the sort proposed by Waldmann and colleagues. With respect to this latter issue, several potentially fruitful avenues for future empirical research come to mind. In particular, as we've just seen, in so far as a Quinn-style formulation of the DDE hypothesis seems most promising, more work needs to be done on cases where it is made more salient whether or not the involvement of the person that suffers harm is causally necessary for achievement of the agent's goal.

References

Anscombe, G.E.M. (1963). *Intention (2nd edition)*. Oxford: Blackwell.

Arbarbanell, L. & Hauser, M. (2010). Mayan morality: An exploration of permissible harms. *Cognition*, 115, 207-224.

Bennett, J. (1995). *The Act Itself*. Oxford: Clarendon Press.

Bratman, M. (1999). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge: Cambridge University Press.

Connell, F.J. & Kaczor, C. (2013). Principle of Double Effect. In R.L. Fastiggi (ed.), *New Catholic Encyclopedia Supplement 2012-13: Ethics and Philosophy*. Farmington Hills, MA: Gale, Cengage Learning.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgement: Testing three principles of harm. *Psychological Science*, 17, 1082-1089.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Oxford University Press.

Davis, N. (1984). The Doctrine of Double Effect: Problems of interpretation. *Pacific Philosophical Quarterly*, 65, 107-123.

Fischer, J.M., Ravizza, M., & Copp, D. (1993). Quinn on Double Effect: The problem of closeness. *Ethics*, 103, 707-725.

Fitzpatrick, W. (2006). The intend/foresee distinction and the problem of "closeness". *Philosophical Studies*, 128, 585-617.

Foot, P. (1967). The problem of abortion and the Doctrine of Double Effect. *Oxford Review*, 5, 5-15.

Greene, J.D. (2013). *Moral Tribes*. New York: Penguin.

Greene, J.D., Cushman, F., Stewart, L., Lowenberg, K., Nystrom, L.E., & Cohen, J.D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111, 364-371.

Hauser, M., Cushman, F., Young, L., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind and Language*, 22, 1-21.

Hills, A. (2007). Intentions, foreseen consequences and the Doctrine of Double Effect. *Philosophical Studies*, 133, 257-283.

Huebner, B., Hauser, M., & Pettit, P. (2011). How the source, inevitability, and means of bringing about harm interact in folk-moral judgments. *Mind and Language*, 26, 210-233.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis,* 63, 190-193.

Liao, S.M., Wiegmann, A., Alexander, J. & Vong, G. (2012). Putting the trolley in order: experimental philosophy and the loop case. *Philosophical Psychology*, 25, 661-671.

Masek, L. (2010). Intentions, motives and the Doctrine of Double Effect. *Philosophical Quarterly*, 60, 567-585.

Mikhail, J. (2000). *Rawls' Linguistic Analogy*. PhD thesis. Cornell University.

Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and The Cognitive Science of Moral and Legal Judgment*. New York: Cambridge University Press.

Prinz, J. (2008). Is morality innate? In W. Sinnott-Armstrong (ed.), *Moral Psychology, Volume 1: The Evolution of Morality: Adaptation and Innateness*. Cambridge, MA: MIT Press.

Quinn, W. (1989). Actions, intentions, and consequences: The Doctrine of Double Effect. *Philosophy and Public Affairs* 18, 334-351.

Schwitzgebel, E. & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind and Language*, 27, 135-153.

Sterelny, K. (2010). Moral nativism: A sceptical response. *Mind and Language*, 25, 279-297.

Thomson, J.J. (1985). The Trolley Problem. *The Yale Law Journal*, 94(6), 1395-1415.

Waldmann, M. & Dieterich, J.H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, 18, 247-253.

Waldmann, M. & Wiegmann, A. (2010). A double causal contrast theory of moral intuitions in trolley dilemmas. In S. Ohlsson & R. Catrambone (eds.), *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society* (2589-2594). Austin, TX: Cognitive Science Society.

Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*, 25, 813-836.

Wiegmann, A., & Waldmann, M. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, 131, 28-43.

Woodward, P.A. (2001). *The Doctrine of Double Effect: Philosophers Debate a Controversial Moral Principle*. Notre Dame, IN: University of Notre Dame Press.