

## DEFENDING DOUBLE EFFECT

*Ralph Wedgwood*

### *Abstract*

This essay defends a version of the Doctrine of Double Effect (DDE) – the doctrine that there is normally a stronger reason against an act that has a bad state of affairs as one of its intended effects than against an otherwise similar act that has that bad state of affairs as an unintended effect. First, a precise account of this version of the DDE is given. Secondly, some suggestions are made about why we should believe the DDE, and about why it is true. Finally, a solution is developed to the so-called ‘closeness problem’ that any version of the DDE must face.

**0.** One form of deontological ethics involves the so-called Doctrine of Double Effect (DDE).<sup>1</sup> As I shall interpret it here, the DDE is the thesis that there is normally a *stronger* reason against an act if that act has a bad state of affairs (like an innocent person’s death) as one of its *intended* effects than if that bad state of affairs is merely one of the act’s *unintended* effects. In short, according to the DDE, it is harder to justify an act that has a bad effect if that effect is intended than if it is not intended.

Many writers have criticized the DDE.<sup>2</sup> In this essay, I shall offer a partial defence, by giving a precise account of one version of the doctrine, and answering some of the objections that have been raised against it.

**1.** The version of the DDE that I shall defend here is not exactly identical to the versions that have been discussed by other philosophers. For example, consider the version of the DDE that is discussed by T. M. Scanlon (2009, 1):

The doctrine of double effect holds that an action that aims at the death of an innocent person, either as its end or as a means to its end, is always wrong.

There are several differences between Scanlon’s version of the doctrine and the version that I shall defend here. First, Scanlon’s version condemns all acts that ‘aim at’ the death of an innocent person, regardless of whether these acts succeed in achieving this aim or not. The version of the DDE that I shall defend here, by contrast, concerns only those acts that *succeed* in realizing the bad state of affairs that they were aiming at. In my view, acts that aim at a bad state of affairs but *fail* to achieve this aim fall into a somewhat different category – which unfortunately I shall not have time to discuss here.

Secondly, the version of the doctrine that Scanlon discusses is narrowly focused on acts that aim at the *death of an innocent person*. But it seems clear that if this version of the doctrine is true, this will not be because of anything special about *deaths* in particular; instead it will be because of a fundamental difference between the roles of intended effects and unintended effects in generating reasons against action. For this reason, I shall focus on a significantly broader version of the doctrine here, which is concerned with all acts that have a *bad state of affairs* among their intended effects. My version can presumably explain the more restricted version that is concerned solely with deaths, since an innocent person’s death is presumably normally a ‘bad state of affairs’ in the relevant sense.

---

1 The doctrine goes back at least as far as Thomas Aquinas’s discussion of self-defence; see *Summa Theologica*, IIa IIae, 64, 7.

2 Notably, the DDE has been criticized by some proponents of deontological ethics, such as Judith Thomson (1991 and 1999) and T. M. Scanlon (2009). I have already responded, at least briefly, to their objections elsewhere; see Wedgwood (2011). This is why I shall focus on different objections here.

As I shall sometimes put it, when one intends a state of affairs that is bad in the relevant way, one's intention is a 'bad intention'; and when one intends a state of affairs that is not bad in this way, one's intention is an innocent or permissible intention. So another way to put the central idea of the DDE is by saying that at least normally, other things being equal, there is a stronger reason against an act that is the successful execution of a bad intention than against an otherwise similar act that is the execution of an innocent or permissible intention.

Finally, there is another crucial difference between Scanlon's version of the DDE and mine. Scanlon's version is *absolutist*: it implies that acts that have an innocent person's death as one of their intended effects are *always* wrong. But Warren Quinn's (1989) version of the doctrine seems more plausible: according to Quinn's version, the fact that an act has a bad state of affairs as one of its intended effects normally *strengthens* the reason *against* the act, but it does not *invariably* make the act *impermissible* (since sufficiently strong countervailing reasons may make the act permissible after all).

The central claim of my version of the DDE, then, is that there is normally a stronger reason against an act if the act has a bad state of affairs as one of its intended effects than if it has that bad state of affairs as one of its unintended effects. Bad intended effects count more strongly against actions than bad unintended effects. Moreover, the difference in strength between these two reasons against acting is non-trivial: at least sometimes this difference can make it the case that an act that is done with a bad intention is impermissible, while an otherwise exactly similar act that is done with an innocent intention is permissible.

There are some further questions about how exactly to interpret the doctrine that need to be addressed before proceeding. First, what is meant here by an 'effect' of an act? I shall use the term broadly here, so that an 'effect' of an act comes to exactly the same thing as a 'consequence' of the act: it is simply a state of affairs that would obtain if the act were to be performed, but might not obtain if the act were not performed.<sup>3</sup>

Secondly, what do I mean by saying that some consequences of an action are 'bad'? The sort of badness that I have in mind here is an *agent-neutral* sort of badness. A state of affairs that is 'bad' in this agent-neutral way is not just *bad for me* (which might be compatible with its being good for everyone else), but bad in a way that makes it appropriate for *everyone* to regret or lament the state of affairs (regardless of who they are, or how precisely they are related to this state of affairs). There is much more that could be said about this idea of the agent-neutral badness of states of affairs; but I shall not pursue these issues here. For the purposes of this discussion, I shall have to rely on an intuitive grasp of this sort of agent-neutral badness. However, I have *not* said that when a state of affairs is bad in the relevant way, that state of affairs is *intrinsically* bad. I have left it open that the states of affairs that are bad in the relevant way may be merely *extrinsically* rather than intrinsically bad. (We shall return to this point when we discuss the 'closeness problem' for the DDE in the last section of this paper.)

Thirdly, we also need to clarify our talk of 'acts' and 'actions'. One crucial distinction is between act-types and act-tokens.<sup>4</sup> An act-token is a particular act that is actually performed by a particular agent at a particular time. By contrast, an act-type can be performed on many different

---

3 Following David Lewis (1973), I assume here that the 'might'-counterfactual 'If it were the case that *p*, it *might not* be the case that *q*' is equivalent to the negation of the 'would'-counterfactual 'If it were the case that *p*, it *would be* the case that *q*'. The consequences of an act need not always be states of affairs that *would not* obtain if the act were not performed. Suppose that there are three available acts, *A*, *B*, and *C*, and a state of affairs *S*, such that *A* and *B* would each result in *S*'s obtaining, while *C* would result in *S*'s not obtaining. Then it would not be true that if you were not to do *A*, *S* would not obtain (since if you did not do *A*, you might do *B* – in which case *S* would still obtain); but *S* surely still counts as a consequence of *A*.

4 This terminology has now become standard, but it was first used in this way by Alvin Goldman (1970), alluding to the well-known distinction between *linguistic* types and tokens.

occasions, and by many different agents. Moreover, an act-type can exist even if it is never actually performed – the most that is required for the existence of an act-type is the *possibility* of there being a performance of that type.

Given my interpretation of the DDE, the kind of judgments that we need to focus on here are judgments about the strength of an agent's reasons for or against various available actions – including actions that the agent never actually performs. It seems that these are most naturally taken as judgments on *act-types*, considered as possible options for a particular agent in a particular choice situation. For example, the statement 'There is a reason for you to call your mother tomorrow' states that there is a certain relation (the reason-for relation) between you, an act-type (calling your mother) and an *occasion* (tomorrow).

Here, however, there is a complication that we need to take account of. Some act-types are more *specific* and *detailed* than others. For example, the act of flying from Mexico to Toronto is a more specific act-type than the act of flying somewhere in North America: it is necessary that anyone who does the first act also does the second, but not *vice versa*. Similarly the acts that are, intuitively, proper parts of other acts – as covering the distance from Mexico to the US border is part of the act of travelling the whole distance from Mexico to Toronto – are also less specific than the acts of which they are proper parts, since here again, it is possible to do the former act without doing the latter but not *vice versa*.

One way in which an act-type can be more specific than another is if the more specific type itself contains a specific intention. Thus, in the famous trolley case that was originally due to Philippa Foot (1978, 23), one relatively general act-type might be *diverting the runaway trolley onto the side track*; but there are also two more specific act-types, each of which incorporates a specific intention with which one might divert the trolley – *diverting the trolley in order to kill the person on the side track*, and *diverting the trolley in order to save the five people on the main track*. I shall call the act-types that do not incorporate the intention in this way the 'thin' act-types, while the act-types that do incorporate the intention in this way will be called the 'thick' act-types.

Fundamentally, as I am interpreting it, the DDE is concerned with the *thick* act-types, such as *diverting the trolley in order to kill the person on the side track*, and *diverting the trolley in order to save the five people on the main track*. The central claim of the DDE, as I shall interpret it, is that there is normally a stronger reason against a thick act-type that involves successfully executing a bad intention (that is, an intention to bring about a state of affairs that is bad in the relevant way) than an otherwise similar act-type that involves executing innocent intentions instead.

Finally, we need to make it clearer what we mean here by an 'intention'. Much of the literature on the DDE is deplorably unclear about this. In particular, it is crucial to distinguish between (i) the intention with which the action is performed, and (ii) the motivation that lies behind and explains the intention. The motivation of an action consists of the whole of the mental process that terminates in the action; this process typically includes many mental states in addition to intentions – such as desires, emotions, wishes, and beliefs (including of course normative and evaluative beliefs). Within this process, the intentions are in a sense the *last* purely mental component immediately preceding the action; and the action itself just *is* the execution of the intention. In this sense, the agent's intention in acting is no mere *antecedent* of the act, but an essential constituent of the act itself.

The crucial feature of an intention, then, which differentiates it from the other mental states that are involved in the motivation of action, is that the act itself *is* the execution of the intention. Executing an intention involves behaving in a way that is *guided* and *controlled* by that intention. Consider, for example, your intention to make an omelette. This intention consists in a conception of a possible causal sequence of events that might take place, one after another (breaking and beating some eggs, melting some butter in a frying pan, and so on), such that this sequence of events will result in your making an omelette. In executing this intention, you will

be constantly monitoring events as they unfold in your kitchen, and constantly adjusting your behaviour in such a way as to ensure that the actual sequence of events conforms to this conception.<sup>5</sup>

Since the fundamental role of an intention is to guide or regulate the agent's voluntary behaviour, the content of an intention always concerns events that the agent takes to be within her power to control. So, for example, if you place a bomb on an aeroplane, in order to destroy some incriminating documents that are on the plane, your intention is to blow up the whole plane and all its contents, since you know perfectly well that this course of action cannot exercise any control over the documents except through acting on the plane and all its contents.

The version of the DDE that I shall defend here concerns intentions in this strict sense of the term. It does *not* concern the agent's motives in general. I shall take no stand on whether all of the agent's motives make a difference to the permissibility of the action. The only doctrine that I shall defend is the thesis that – whatever the agent's other motives may be – there is a stronger reason against an act that has a bad state of affairs as one of its *intended* effects than against an otherwise similar act that has that bad state of affairs as one of its *unintended* effects.

Even though I have in this way restricted the version of the DDE that I am considering, so that it concerns only intentions in this strict sense of the term, it is still a strong and controversial doctrine. First, I have generalized the doctrine so that it covers *all* cases in which a bad state of affairs is intended, not just intentional killings. Secondly, the doctrine implies that intentions have *intrinsic* or *non-derivative* ethical significance. There is a stronger reason against bringing about bad effects through executing an intention to do so than against bringing about such bad effects without intending to precisely *because* of the difference that consists in the presence or absence of this intention. The greater strength of the reason against acting is not explained by something that is merely correlated with the intention, but by the intention itself.

So far, I have focused on the thick act-types, like *diverting the trolley in order to kill the person on the side track*, and *diverting the trolley in order to save the five people on the main track*. As I have explained, the DDE implies that there could be cases in which the first of these two thick act-types is impermissible, while the second thick act-type is permissible. That is, there could be cases in which it is impermissible to divert the trolley in order to kill the person on the side track, but permissible to divert the trolley in order to save the five people on the main track; indeed, in some cases of this sort, it might even be true that you *ought* to divert the trolley onto the side track in order to save the five on the main track (although of course you ought not to do it in order to kill the one person on the side track).

However, what should the DDE say about the *thin* act-types, like simply *diverting the trolley*? Presumably, philosophers who follow the 'actualist' view of Frank Jackson and Robert Pargetter (1986) will assume that proponents of the DDE must say that a thin act-type is permissible only if the agent *would* do it with a permissible intention *if he did it*.<sup>6</sup> But is this assumption obviously correct? Suppose that if you were to divert the trolley, you would do it with the bad intention of killing the man on the side track. Must the proponents of the DDE say that in that case you ought not to divert the trolley at all?

It seems to me that the proponents of the DDE should not say this. If it is only because of your utter wickedness that you are such that if you diverted the trolley, you would do it with this bad intention, then the proponents of the DDE should *deny* that this entails that you ought not to divert the trolley. After all, you *ought not* to be such that, were you to divert the trolley, you would do it with this bad intention; you ought to be such that, were you to divert the trolley, you

---

5 For some particularly illuminating discussions of intention, see Bratman (1987) and Mele (2000).

6 Judith Thomson (1991, 293) seems to make this assumption, since she claims that the DDE entails that a would-be bomber would have to 'decide whether he may drop the bombs by looking inward for the intention with which he would be dropping them if he dropped them'. As I explain here, I believe that it is quite mistaken to interpret the DDE as committed to this.

would do it with a permissible intention. So it could even be true that you *ought* to divert the trolley – although of course you should only divert the trolley with the good intention of saving the five on the main track (not with bad intention of killing the one person on the side track). In general, proponents of the DDE can happily accept that the fact that an agent is such that, if he were to do a certain thin act-type, he would do it with a bad intention does not entail that it is impermissible for the agent to do the thin act-type; indeed, this fact does not even entail that it is not true that the agent *ought* to do the thin act-type.

Instead, it seems to me, the natural way to extend the DDE from thick act-types to thin act-types is by invoking the following two principles. First, a thin act-type is *permissible* if and only if there is *some* permissible thick act-type of which that thin act-type is a part. Secondly, a thin act-type is *impermissible* if and only if *every* available thick act-type of which that thin act-type is a part is impermissible. If diverting the trolley is part of a permissible thick act-type (such as diverting the trolley in order to save the five people on the main track), it follows that it is permissible to divert the trolley – although if there is a danger that the agent will divert the trolley with a bad intention, it may be misleading just to assert that he may divert the trolley, without adding that he may only divert the trolley with this permissible intention, not with the bad intention of killing the person on the side track.

This, then, is the version of the DDE that I shall defend here. In the next section, I shall touch briefly on the question of why we should think that anything like this version of the DDE is true.

2. The DDE is sometimes viewed as if it is accepted only by a few Catholic moral theologians, who have been influenced by Thomas Aquinas. But this view seems quite mistaken to me. The fundamental idea behind the DDE is deeply engrained in a great many traditions of serious moral thinking – including legal and philosophical thinking, as well as religious thinking. Thus, for example, in the law, this idea reappears as the distinction between (i) a ‘direct’ or ‘purposeful’ intention, and (ii) an ‘oblique intention’ (as is often confusingly called), which consists simply of the effects that the agent *knew* the act to be likely to produce. This distinction is relevant to the law in two main ways. First, there are several criminal offences that essentially involve a direct or purposeful intention (as opposed to a mere oblique intention); for example, this distinction plays a fundamental role in the laws regarding the conduct of military personnel in warfare. Secondly, even with criminal offences that do not require a direct intention, such a direct intention will normally count as an aggravating circumstance, increasing the gravity of the offence.<sup>7</sup>

So I believe that the DDE is presupposed by a large number of moral beliefs that have been accepted by a large and diverse collection of moral thinkers, over a long period of time. Our default assumption should be that any idea that has been presupposed by a wide moral range of moral beliefs that have been comparatively stable over time and across cultures is likely to contain some truth buried inside it somewhere.

There is also another reason that many people have for accepting this doctrine – although unfortunately I shall not be able to set out this reason in detail here. Many of us have clear intuitions about a certain range of cases, and the DDE seems to provide the best explanation of these intuitions. Some of the philosophers who reject the DDE – such as Scanlon (2009) – have tried to provide alternative explanations for our intuitions about these cases, while philosophers who support the DDE – such as McMahan (2009) – have tried to cast doubt on these alternative explanations. In this way, these cases have already been extensively discussed by a large number of other philosophers. To save time, I shall avoid getting into this dispute here (although it does seem to me that the defenders of the DDE have had stronger arguments). Instead, I shall make a quick suggestion about what might explain *why* the DDE is true.

---

<sup>7</sup> The literature on this topic within legal scholarship is vast. See, for example, Douglas Husak (2009).

One notable explanation of the DDE is that of Thomas Nagel (1986, 181). According to Nagel's explanation, what is especially bad about executing an intention to bring about a bad state of affairs is that in such cases, your 'will' is being 'guided by evil'.<sup>8</sup> My explanation of the DDE is broadly similar to Nagel's, except that I see the DDE as ultimately just an instance of a larger phenomenon.

According to my explanation, whenever the consequences of an act include a bad state of affairs, this grounds a reason against the action; but the strength of the reason does not depend purely on the badness of this state of affairs, but also on what I have called the agent's 'degree of agential involvement' in bringing about this state of affairs. Broadly speaking, I suggest that there are two dimensions along which one can be to a greater or lesser degree 'agentially involved' in bringing about a state of affairs: the first dimension is *causal*; the second dimension is *intentional*. Along the causal dimension, there is a crucial difference between *actively causing* a state of affairs and merely *failing to prevent* that state of affairs (in effect, this is what many philosophers think of as the distinction between *doing* and *allowing*). If you merely *fail to prevent* a state of affairs from coming about, then your degree of agential involvement in bringing about that state of affairs is much less than if you *actively cause* that state of affairs to come about.

In addition to this causal dimension of agential involvement, there is also the *intentional* dimension. Other things equal, your degree of agential involvement in bringing about a state of affairs is greater if you directly *intend* that state of affairs than if you do *not* intend that state of affairs – even if you foresaw that your act was likely to result in that state of affairs. Your agency is more involved with a consequence of your act that you intended than with a consequence that you did not intend.

When your act has a bad consequence, the more agentially involved you are in bringing about that consequence, the stronger the reason against the act will be.<sup>9</sup> The bad consequence is more intimately connected to the act, and so the badness of the consequence is more strongly reflected in the reason against the act. In general, the strength of the reason against the act seems to correspond to the weighted sum of the degrees of badness of each of the act's consequences – where the degree of badness of each consequence is weighted by the degree of agential involvement that the agent has in that consequence.

This then is my explanation of why the DDE is true. According to this explanation, the DDE flows from a completely general feature of reasons for action: namely, in the significance for reasons for action of the agent's degree of agential involvement in the consequences of the act. So the DDE is not just a widely-accepted idea that is supported by a wide range of intuitions; it can also be explained as flowing from a pervasive and fundamental feature of the normative domain.

3. Finally, I should like to address what has come to be known as the 'closeness problem'.<sup>10</sup> For example, consider one of the cases of Judith Thomson (1985). A runaway trolley is hurtling towards five people who are trapped on the railway track. In this case there is no side track.

---

8 Of course, it may be that your will is not *wholly* 'guided by evil', since even if your intended means are bad, your ultimate end may be good, and not bad at all. Still, your will is guided by the *whole* of your intention – including your intention to use the bad means, as well as by your intention to achieve the good end. So your will is at least *partially* guided by evil in these cases.

9 Similarly, I believe, if your act has a *good* consequence, the more agentially involved you are in bringing about that consequence, the stronger the reason *in favour* of that act will be. For more discussion of this idea of 'agential involvement', see Wedgwood (2009).

10 For discussions of the 'closeness problem', see Hart (1968), Quinn (1989), Bennett (1995, Chap. 11), and Predelli (2004).

Instead, a large man carrying a heavy back-pack is standing on a footbridge that crosses the railway line, and if you push the man off the bridge, the trolley will collide with him and grind to a halt before it can hit the five. Many advocates of the DDE assumed that one reason why pushing the man off the bridge is objectionable is because in so doing you intend his death.

On further reflection, however, it seems that if you push the man off the bridge, you could legitimately claim *not* to intend the man's death. You only intend that he should *collide* with the trolley, so that the collision will bring the trolley to a halt. If by some miracle he survives the collision, your intentions would not require your doing anything else that would ensure his death; indeed, it would be entirely compatible with your plan if you also did everything possible to maximize the chances that he survives the collision. Some advocates of the DDE have responded to this point by making the following move: they have suggested that the man's collision with the trolley is sufficiently 'close' to the man's death that we can legitimately 'redescribe' your intending the collision as tantamount to your intending his death.<sup>11</sup>

It seems to me, however, that this is a fatal move for proponents of the DDE to make. If the content of the intentions with which an agent is acting is not an objective psychological truth about the agent, then it is radically unclear how intentions could have the ethical significance that the DDE takes them to have. But if it is an objective psychological truth what the contents of your intentions are, then we cannot simply 'redescribe' your intentions in whatever way seems convenient to us.

As we have seen, the intention with which you act is a thought that in the relevant way *guides* or *regulates* your behaviour in acting. When you push the man off the bridge, the thought that is guiding you is 'Let's make sure that he collides with the trolley' – not 'Let's make sure that he gets killed'. For these reasons, the proponents of the DDE must accept that in the bridge case, you intend the collision but *not* the death.

Fortunately, this is no problem for my version of the DDE, since my version is not restricted to cases of intending *death*, but applies to all cases where one intends a *bad state of affairs*. So long as the collision counts as a relevantly 'bad state of affairs', intending the collision will be in my sense a bad intention, and my version of the DDE will still apply to this case. Indeed, it seems that your intention would be even *worse* if you intended not just the collision but the death as well; this seems to be because in the relevant sense the death is itself a *worse* state of affairs than the collision.

So to solve the closeness problem, we need to find a conception of a 'bad state of affairs' according to which the man's colliding with the trolley is a bad state of affairs, but the man's death is an even worse state of affairs.

It might seem obvious: surely the collision is a bad state of affairs simply because it causes the man to suffer harm? But it is equally true that the trolley's being diverted onto the side track in the original trolley case causes harm to the person on the side track. If the trolley's being diverted onto the side track counts as a relevantly 'bad state of affairs', then intending to divert the trolley onto the side track will also be a bad intention, and so the DDE will apply to the original trolley case as well.

In short, to solve this problem we need a conception of 'bad states of affairs' that has the following implications: the man's death and the man's colliding with the trolley are both bad states of affairs, although the death is a worse state of affairs than the collision; and the trolley's being diverted onto the side track is not in the relevant sense a bad state of affairs at all.

It is clear that the man's colliding with the trolley is not an *intrinsically* bad state of affairs. It is bad only because of certain additional attendant circumstances (for example, it would not be bad if the man were completely invulnerable to any injuries that the collision might

---

<sup>11</sup> Alison MacIntyre (2001) assumes that proponents of the DDE will have to use the notion of an 'intention' in this way. Fortunately for me, most of MacIntyre's criticisms do not apply to versions of the DDE that follow my recommendation and avoid using the notion of an 'intention' in this way.

cause). But in fact, the man's death is also not intrinsically bad either. The man's death will certainly be a bad state of affairs if the man has a good life and wishes to go on living. But if he is suffering from an excruciating degenerative disease and longs for death, then his death is arguably not a bad state of affairs. So the man's death is also not intrinsically bad: it is bad only because of certain additional attendant circumstances (such as the fact that the man wishes to go on living, or that his death deprives him of a reasonably good life, or the like).

As I am thinking of intrinsic value, a state of affairs has a certain degree of intrinsic value if and only if it is metaphysically necessary that that state of affairs must have that degree of intrinsic value in any possible world in which it exists. It follows that the only states of affairs that have intrinsic value are tremendously *detailed* states of affairs: each such state of affairs includes within itself everything that is relevant to determining the degree of intrinsic goodness or badness that it has.<sup>12</sup> It is only such highly detailed states of affairs that are, as Zimmerman (2001, 142) would put it, 'evaluatively adequate'.

The states of affairs that we *intend* are almost never such highly detailed states of affairs. Even if a murderer intends his victim's death, he will typically not intend in addition that his victim should be deprived of a life that is worth living. Since we are looking for a kind of 'badness' that is exemplified by the states of affairs that the agent *intends*, it seems that it will have to be some sort of *extrinsic* badness.

Here is a proposal about what the relevant sort of extrinsic badness is. Suppose that you are a reasonably virtuous person, and you hear the news that someone has collided with a runaway trolley. You would presumably respond by thinking, 'Oh no, how awful! That sounds terrible!' Now suppose that you hear the news that a runaway railway trolley was diverted onto a side track. You would naturally respond by thinking, 'So what? That doesn't sound very interesting.' In short, a person's colliding with a fast-moving railway trolley is in some way *bad news*, while a trolley's being diverted onto a side track is not bad news in the same way.

When you take the person's colliding with the trolley to be bad news, it seems that you would somehow be being guided by your knowledge of a range of *ceteris paribus* moral generalizations – such as, for example, the generalization that other things equal, and under normal conditions, when a collision between a person and a runaway railway trolley occurs, the person suffers serious injury or even death as a result. In general, to be a virtuous agent, it is not enough just to know the pure moral truths (of the sort that might form the fundamental principles of an abstract moral theory); it is also crucial to know a range of true *ceteris paribus* moral generalizations of this sort. A virtuous agent will form *expectations* on the basis of these *ceteris paribus* generalizations; it is these expectations that the virtuous agent will express by greeting some pieces of information as good news, and others as bad news.<sup>13</sup>

Roughly, then, I propose that for a state of affairs *S* to count as a bad state of affairs in the relevant way is for the following two conditions to be met: first, a virtuous agent, guided by her knowledge of these true *ceteris paribus* moral generalizations, would form the kind of expectations about *S* that would lead her to view it as bad news in this way; and secondly, in this particular case, these expectations are borne out – that is, things turn out badly in more or less the very way in which *S* would lead such a virtuous agent to expect them to.

This is clearly a kind of extrinsic badness, since the badness of a state of affairs can vary from possible world to possible world as the true *ceteris paribus* generalizations also vary from

---

12 For more discussion of this aspect of intrinsic value, see Wedgwood (2009) and Zimmerman (2001).

13 If pushing the man off the bridge is rational, it is presumably done with the intention of solving a certain practical problem – the problem of how to save the five from the runaway trolley. This problem only exists because the trolley is moving sufficiently fast to pose a lethal risk to the five. So it seems to me that in pushing the man off the bridge, you are acting with the intention of ensuring that he collides with (and so halts) a trolley that is moving at that sort of life-threatening speed. It is surely especially clear this intended state of affairs is a bad state of affairs in this sense.

world to world. Nonetheless, it is still a kind of *agent-neutral* badness: all virtuous agents will view the state of affairs as bad news in the same way, regardless of their particular relationships to the individual people or objects that are involved in that state of affairs.

Whenever *S* counts as a bad state of affairs of this kind, and an agent intends this state of affairs *S*, this intention counts as a ‘bad intention’ in my sense. By intending this state of affairs *S*, the agent’s will is being ‘guided’ by a certain kind of ‘evil’ – that is, by what virtuous agents would regard as bad news; in this way, in effect, the agent’s will is swimming against the normative tide that is created by these *ceteris paribus* generalizations.

According to my proposed interpretation of the DDE, then, when an agent successfully executes a bad intention of this sort, there is a stronger reason against the act than there would have been against an otherwise similar act that is not done with such a bad intention.

Let us examine how this solution to the closeness problem will deal with some problem cases, starting with a case that was first discussed by Foot (1978). Suppose that you blow up a man who is trapped in the mouth of a cave, so that you and your friends can escape from the cave before being drowned by the rising tide. In this case, you intend that the man should be blown to pieces, but not that he should die: the thought that is guiding your behaviour is ‘Let’s ensure that he is blown to pieces’ – not ‘Let’s make sure that he ends up dead’. Still, his being blown to pieces is a bad state of affairs in the relevant way, and so the intention with which you act is still a bad intention.

Now consider the following case, which is due to Stefano Predelli (2004). Suppose that the Reds are at war with the Blues, and the Red military command wishes to induce the Blue government to surrender immediately, by making them believe that the entire population of the Blues’ second-largest city has been annihilated. All communications between the Blue capital and the Blues’ second city have been severed by bombing, and so the only way for the Reds to convince the Blue government that the second city has been annihilated is by detonating bombs in the air above the city where they can be seen from the hills around the Blue capital. Unfortunately, detonating these bombs will have the foreseen but unintended effect of annihilating the city. In this case, the detonation of the bombs is intended, but the annihilation of the city is not. Nonetheless the detonation of the bombs is surely still bad news in the relevant way. So intending the detonation of the bombs is a bad intention in my sense.

Now, let us consider some cases to which my interpretation of the DDE does not apply. First, suppose that I intend to catch your attention, so that you will look up and notice that I am in the room, because this is the agreed signal that will alert a spy who is observing us, thereby enabling the spy to defuse a bomb that would otherwise kill many people. (Suppose that I know that there is no other way in which the bomb can be defused if I do not catch your attention in this way.) I have also just found out that you have a bizarre condition so that when I catch your attention in this way, it will result in your instant death.

In this case, my intention is not a bad intention: there is no true *ceteris paribus* moral generalization to the effect that under normal conditions, catching a person’s attention results in any serious injury or harm. So my interpretation of the DDE does not apply here: there is presumably a reason against the act grounded in the fact that the act will kill you, but there is no reason (of the kind distinctive of the DDE) grounded in the fact that this act is the execution of a bad intention. In this case, Warren Quinn’s (1989) version of the DDE yields a different verdict: in this case, I do intend to involve you in my actions in a certain way, at the same time as knowing that your being involved in that way will harm you – and that is enough for the case to involve Quinn’s interpretation of the DDE. It seems to me that my verdict on the case is more plausible than Quinn’s.

Finally, suppose that a judge intends to send a convicted offender to prison. Admittedly, there *is* a true *ceteris paribus* generalization to the effect that, other things equal, when an offender is imprisoned, the offender’s family suffers. But there are also other true *ceteris paribus* generalizations here as well – when a convicted offender is imprisoned, the offender’s victims

receive justice, other criminals are deterred from committing crimes, and so on. On balance, then, a virtuous agent would not react to the news that a convicted criminal has been imprisoned by thinking ‘Oh no! That sounds terrible!’ So the intention to send the offender to prison is not obviously a bad intention in the relevant way.

I suggested above that an intention to bring about a man’s death would typically be *worse* than an intention to bring it about that the man collides with the trolley. Presumably, this is because the man’s death is in the relevant way a worse state of affairs than the collision. In general, if one state of affairs  $S_1$  is in the relevant way *worse* than a second state of affairs  $S_2$ , then an intention to bring about  $S_1$  will typically be a worse intention than an intention to bring about  $S_2$ .

How are we to make sense of the relevant notion of ‘worse’ states of affairs? The *ceteris paribus* moral generalizations that I referred to above are in effect generalizations over possible cases. This suggests that it may be possible, at least in principle, to develop a *measure* on this space of possibilities – a measure that would presumably have the structure of a *probability function*. (Intuitively, this probability function would correspond to the degrees of belief that ideally rational agents would have if their knowledge consisted *solely* of all and only these true *ceteris paribus* moral generalizations, and did not include any knowledge about the particular case at hand.) Then we can use this probability function to define the *expected badness* of a state of affairs.

Specifically, this is what I have in mind. Let  $P_{CP}$  be the probability function that corresponds to these *ceteris paribus* moral generalizations in the way that I have described. If  $S$  is a state of affairs, and  $W$  is a possible world, let  $IV(W)$  be the degree of intrinsic value of  $W$ . Then we can say that if  $S$  is in the sense that I have just defined a bad state of affairs, its precise degree of badness is the conditionally expected value of  $IV(W)$  on the assumption of  $S$ , according to the probability function  $P_{CP}$  – that is,

$$\sum_W P_{CP}(W|S) IV(W).$$

In other words,  $S$ ’s degree of badness is the weighted sum of the degrees of intrinsic value of each relevant possible world  $W$ , weighted by the conditional probability (according to this probability function  $P_{CP}$ ) of  $W$  on the condition that  $S$  occurs.

This conception of a state of affairs’ degree of badness can explain why the man’s death is a worse state of affairs than the collision between the man and the trolley. Although the man’s collision with the trolley is bad news, the man’s death is even worse news: not all collisions result in death, and so the man’s death has an even greater expectation of harm than the collision. For this reason, intending the death is an even worse intention than merely intending the collision.

We can put this point in terms of the notion of ‘agential involvement’ that I introduced in Section 2. If your act has an intrinsically bad consequence – like the man’s losing a life worth living through being killed by the collision with the trolley – and you intend some state of affairs that is part of that consequence, then the worse this intention is (in the sense that I have just defined), the greater your degree of agential involvement in this bad consequence, and the stronger the reason against the act.

In this way, it seems possible to defend a version of the DDE against the closeness problem. In general, I hope to have made it plausible here that the DDE has the resources to rebut many of the criticisms that have been raised against it.<sup>14</sup>

---

14 I am grateful to audiences at the University of Toronto, Trinity College Dublin, the Australian National University, and a conference held in memory of Philippa Foot at Somerville College, Oxford, and to several Oxford colleagues of mine, for helpful comments. I wrote this paper during my tenure of a Research Fellowship from the Leverhulme Trust, to whom I also express my gratitude.

Merton College  
Oxford OX1 4JD  
UK

ralph.wedgwood@merton.ox.ac.uk

## References

- Bennett, Jonathan (1995). *The Act Itself* (Oxford: Clarendon Press).
- Bratman, Michael E. (1987). *Intentions, Plans, and Practical Reason* (Cambridge, Massachusetts: Harvard University Press).
- Foot, Philippa (1978). 'The Problem of Abortion and the Doctrine of the Double Effect', reprinted in Foot, *Virtues and Vices* (Oxford: Blackwell).
- Goldman, Alvin (1970). *A Theory of Human Action* (Englewood Cliffs, New Jersey: Prentice-Hall).
- Hart, H. L. A. (1968). 'Intention and Punishment', reprinted in Hart, *Punishment and Responsibility* (Oxford: Oxford University Press).
- Husak, Douglas (2009). 'The Costs to Criminal Theory of Supposing that Intentions are Irrelevant to Permissibility' *Criminal Law and Philosophy* 3: 51–70.
- Jackson, Frank, and Pargetter, Robert (1986). 'Oughts, Options, and Actualism', *Philosophical Review* 95: 233–55.
- Lewis, David (1973). *Counterfactuals* (Oxford: Blackwell).
- MacIntyre, Alison (2001). 'Doing Away with *Double Effect*', *Ethics*, 111(2): 219–255.
- McMahan, Jeff (2009). 'Intention, Permissibility, Terrorism, and War', *Philosophical Perspectives* 23: 345–72.
- Mele, Al (2000). 'Goal-Directed Action: Teleological Explanations, Causal Theories, And Deviance', *Philosophical Perspectives* 14: 279–300.
- Nagel, Thomas (1986). *The View from Nowhere* (Oxford: Clarendon Press).
- Predelli, Stefano (2004). 'Bombers: Some Comments on Double Effect and Harmful Involvement', *Journal of Military Ethics* 3: 16-26.
- Quinn, Warren (1989). 'Actions, Intentions, and Consequences: The Doctrine of Double Effect', *Philosophy & Public Affairs* 18 (1989): 334-351.
- Scanlon, T. M. (2009). *Moral Dimensions: Permissibility, Meaning, Blame* (Cambridge, Massachusetts: Harvard University Press).
- Thomson, Judith (1985). 'The Trolley Problem', *Yale Law Journal* 94: 1395-1415.
- Thomson, Judith (1991). 'Self-Defense', *Philosophy and Public Affairs* 20: 283–310.
- Thomson, Judith (1999). 'Physician-Assisted Suicide: Two Moral Arguments', *Ethics* 109: 497–518.
- Wedgwood, Ralph (2009). 'Intrinsic Values and Reasons for Action', *Philosophical Issues* 19: 342–363.
- Wedgwood, Ralph (2011). 'Scanlon on Double Effect', *Philosophy and Phenomenological Research*, forthcoming.
- Zimmerman, Michael J. (2001). *The Nature of Intrinsic Value* (Lanham, Maryland: Rowman & Littlefield).