NOÛS 36:4 (2002) 622–642

*Intending Harm, Foreseeing Harm, and Failures of the Will**

DAVID MCCARTHY
University of Bristol

Very many moral theorists accept that what makes a particular morality the correct or true morality is some kind of appeal to what is good and bad for people. And to many people, this suggests a picture of morality as having two quite distinct levels. The most basic level, first order morality, is an account of which behavior is permissible or impermissible. And the idea is that what makes a particular account of the content of first order morality correct is that it is the set of propositions about the permissibility and impermissibility of behavior which, if generally followed or recognized, would result in things that are good and bad for people meeting some appropriate distributive criterion. Something like this is accepted in different forms by writers in both the utilitarian and the contractualist traditions.

On this picture, an agent's reasons for action are not relevant in any very interesting way to the permissibility or impermissibility of her behavior.[1] Very roughly, it is what people do that affects us for better or worse, not what they are thinking while they are doing it. But we do take a strong interest in people's reasons for action, and some of this interest is moral. For example, a student may resent a teacher whose reason for giving her a bad grade was just that the teacher disliked her even though the bad grade was the right grade. So as well as first order morality, which is concerned with the permissibility and impermissibility of behavior, there is second order morality, which is concerned with various sorts of moral evaluations of agents.[2] Reasons for action have little to do with first order moral evaluation, but are central to second order moral evaluation.

As I said, many moral theorists accept this picture of first and second order morality and think that nothing but confusion comes of forgetting the distinction. But many moral theorists reject it and think that an agent's reasons for

action play an important role in determining whether her behavior is permissible. The best known example of this is the principle of double effect (pde). One version of the pde says that it is always impermissible for an agent to bring about intended harm, but it can be permissible for an agent to bring about merely foreseen harm. Whether it is permissible will depend on what benefits the agent also brings about. Other versions say that bringing about intended harm always constitutes a wrongmaking feature of action over and above the wrongmaking feature which bringing about merely foreseen harm constitutes, but do not go as far as saying that bringing about intended harm is always impermissible.[3] And it is also possible to extend the basic idea of the pde to cover other wrongmaking features of action such as promise breaking,[4] and perhaps allowing harm.[5] But all versions make an agent's intentions play an important role in determining whether her behavior is permissible, and thus reject the first order/second order picture just sketched.

The main reason defenders of the pde offer for accepting it comes from judgments about cases. We are offered pairs of cases which differ only, or almost only, in terms of the agent's reasons for action, and many of us judge that there is something seriously morally amiss in one case but not the other. Taking this judgment on trust, defenders of the pde then invite us to accept that what best explains the judgment is the presence of intended harm in only one of the cases, and that bringing about intended harm constitutes a distinctive wrongmaking feature of action. Defenders of the pde have also given case-independent arguments for accepting the pde, but I will argue that they have not succeeded in giving anyone who accepts the first order/second order picture to start with any reason to accept the pde.

This does not leave us in a comfortable position. It is awkward to accept the pde without having a theoretical rationale for it since the first order/second order picture has initial plausibility. But it is also awkward to reject the pde without making sense of the judgments about cases since few people give no weight to judgments about cases. After having argued against the existing theoretical arguments for the pde, my main goal will be to show that an alternative principle, what I call the mismatch principle (mmp), makes better sense of judgments about cases than the pde, respects the first order/second order picture, and unlike the pde can be given a theoretical rationale. To put it in somewhat old fashioned terms, the cases defenders of the pde offer as support for the pde all involve a distinctive kind of moral failure of the will, but what this moral failure is has little conceptual connection with intending harm. Or so I will argue.

One caveat: although most writers who have defended the pde have seen its defense as central to the defense of commonsense morality, it is very hard to believe that the pde could be responsible for all the major structural features of commonsense morality. For example, commonsense morality draws an important distinction between harming and allowing harm. Very roughly, it accepts the principle of making and allowing (pma), which says, to a first

approximation, that there are strong constraints on harming and only weak constraints on allowing harm. But the pde is a story about the significance of bringing about intended harm, and surely nothing in that will show that it has to treat bringing about foreseen harm and allowing foreseen harm differently. Hence the pde will be a part of a defensible moral theory which captures the major structural features of commonsense morality only if that moral theory contains an independently motivated defense of the pma.[6]

The same is true, I believe, of a number of other structural features of commonsense morality. If these structural features are not defensible, the general project defenders of the pde are committed to will fail, so since we are in the business of criticizing the pde, let us assume that they are defensible. I will therefore make the assumption that there is a defensible account of much of the structure of commonsense morality, including the pma, which respects the first order/second order picture. For various sorts of outcomes, this account will say that bringing about such outcomes is a wrongmaking feature of action. The pde then extends this account by saying that if an agent's bringing about of such outcomes is intended, the degree of wrongmakingness is magnified,[7] thereby breaking the first order/second order picture. I will also use the assumption to show how the mmp makes better sense of the judgments about cases that have been offered as support for the pde while respecting the first order/second order picture. I believe that the assumption is defensible, but I can do nothing to argue this here.[8]

Finally, there has been much discussion of whether the pde can be given the resources to draw the distinctions it needs to draw between intended harm and merely foreseen harm. In my view, the pde is in trouble here. I believe that there is no plausible account of intended harm which will enable the pde to capture all the cases that have been offered by its defenders as support.[9] But this needs extended argument, so here I will give the pde the benefit of the doubt.[10]

Theoretical Arguments in Favor of the PDE

It will help to focus these arguments around a simple pair of cases.

> *Bystander 1* A child is drowning. The only way the agent can save the child is by throwing it a life-preserver. For the life-preserver to reach the child, the agent will have to throw the life-preserver hard enough that it hits the bystander on her backswing, causing the bystander harm H.

> *Bystander 2* A child is drowning. The only way the agent can save the child is by throwing it a life-preserver. Because of where the child is, the agent cannot throw the child the life-preserver directly. She can only get the life-preserver to the child by bouncing the life-preserver off the bystander, causing the bystander harm H.

In each case, the agent asks herself whether her proposed course of action is permissible. She accepts that the correct account of the permissibility and impermissibility of behavior respects the first order/second order picture, and she assumes, as we are assuming, that this account includes the pma. She therefore decides that whether it is permissible to harm the bystander depends on whether the benefit to the child of having its life saved sufficiently outweighs H, and concludes that it does if and only if H is less than a certain critical magnitude. She asks herself how serious H would be, being concerned, for example, that hitting the bystander would knock him in the water, making him drown, which would clearly exceed the critical magnitude. She correctly judges that H is less than the magnitude, so on that proviso, she throws the life-preserver, hitting the bystander on the backswing in B1 and bouncing the life-preserver off the bystander in B2, in both cases making the bystander suffer harm H. I will take it as obvious that in B1, the agent does not intend to hit the bystander, but that in B2 she does. With this pair in mind, let us look at the defenses that have been offered of the pde.

Foot, in discussing a case where the agent's allowing someone to die was (she claims) intended, says that it would be "morally objectionable" for a spectator to be glad that the agent so intended (Foot 1985, 25–26). We can try to apply this idea to the agent rather than the spectator. We might try to say that in B2 the agent is glad of the harm to the bystander, but in B1 is not, and that it is morally objectionable to be glad of harm to another. But what sense of "glad" would make that true?

Here is one sense. In both B1 and B2, if the agent sees the life-preserver in midair and sees that she has not hit bystander, then given her beliefs about the world, this means that she has failed to get the life-preserver to the child, and she will be disappointed as the child will drown. In this sense, the agent is glad to hit the child in both B1 and B2. Here is second sense. In both B1 and B2, had the agent been able to get the life-preserver to the child without hitting the bystander, then she would have done. And in both cases, if she discovers that by some fluke, she has gotten the life-preserver to the child without hitting the bystander, she will be pleased. In this sense, the agent is glad to hit the child in neither B1 nor B2. I can't see any other senses, so I conclude that Foot's idea does not separate B2 from B1.[11]

Quinn argues that various cases motivate a version of the pde which is somewhat different from the traditional version.[12] I will not discuss the differences, but will just examine what Quinn has to say in defense of his version which quite straightforwardly finds a distinctive wrongmaking feature in B2. His first claim is that the pde

> rests on the strong presumption that those who can usefully be involved in the promotion of a goal only at the cost of something protected by their independent moral rights... ought, *prima facie*, to serve the goal only voluntarily. (Quinn 1989, 349)

But so far nothing has been said to distinguish this from cases involving merely foreseen harm: it would be decent for the agent, if there were time, to seek the bystander's consent in B2, but it would also be decent in B1.

Quinn goes on to claim that in the cases in which the pde finds a distinctive wrongmaking feature, the victim is used; not so in the other cases. In one sense, Quinn's claim is true: there seems to be a strong enough counterfactual dependency between the bystander being hit and the child getting the life-preserver in B2, along with the agent's commitment to exploiting this dependency to achieve her goal of the child getting the life-preserver, to say that the agent uses the bystander in B2, but not so in B1. But this is just another way of saying that the agent intended the harm in B2, but didn't in B1, so this sense of "used" will provide no independent support for the pde.

In another sense of "used", if it were true that the agent used the bystander in B2 but not B1, it would follow, on some views, that B2 contains a distinctive wrongmaking feature of action. This is the sense in which there is often said to be a Kantian injunction against using people. This may be what Quinn has in mind: with little explanation, he says that the pde "reflects a Kantian ideal of human community" (Quinn 1989, 350). Without asking whether using people in the Kantian sense *is* a distinctive wrongmaking feature of action, let us ask whether the agent uses the bystander in this sense in and only in B2. Standard accounts say that an agent uses a victim in the Kantian sense if and only if the victim is some way unable to assent to what agent does or is trying to do. But what way of being unable to assent captures the Kantian idea? A first proposal is that the victim is unable to assent if and only if the victim does not assent. But this does not distinguish B2 from B1; there is assent in neither. A second proposal is that the victim is unable to assent if and only if had the victim refused to assent (or had assented), the agent would have behaved no differently, disregarding the victim's refusal. In B1 and B2, the bystander neither assented nor refused to assent, and I did not say whether the agent would have disregarded the refusal had the bystander refused. But I could have added the same story to both B1 and B2, so this proposal would not distinguish B2 from B1. A third proposal is that the victim is unable to assent if and only if what agent is trying to do cannot be achieved with the victim's assent. I believe that this proposal captures the Kantian account; the standard example is deception: if I am trying to deceive you, I cannot achieve this with your assent (Korsgaard 1986). It is also what Quinn seems to have in mind, for immediately after introducing the idea of a Kantian ideal of community, he says: "Each person is to be treated, so far as possible, as existing only for purposes that he can share" (Quinn 1989, 350). But there is no reason at all why the bystander cannot share the agent's purposes in both B1 and B2; nothing about what agent is trying to do in either case would make achieving this impossible with the bystander's assent. The Kantian sense of using people does not distinguish B2 from B1, and therefore does not provide the basis for a defense of the pde.

Nagel begins his defense of the pde with an image: it's as if the agent's intentions magnify, from her perspective, the badness of the harm. When magnified, its badness towers over the good thereby being produced. But as he says, this is just an image, not a justification, and he tries to offer one. Claiming that to intend something is to be guided by it, he says:

> But the essence of evil is that it should *repel* us. If something is evil, our actions should be guided, if they are guided by it at all, towards its elimination rather than towards its maintenance. That is what evil *means*. So when we aim at evil we are swimming head-on against the normative current. Our action is guided by the goal at every point in the direction diametrically opposite to that in which the value of that goal points. … [F]rom the point of view of the agent, this produces an acute sense of moral dislocation. (Nagel 1986, 182)

What does "evil" mean? In one sense, an evil is a serious wrongdoing, but this would make this passage question-begging and is clearly not what Nagel intends (he later endorses a nonabsolute version of the pde). In the other sense, an evil is just a serious harm, and I will take the liberty of rephrasing Nagel's remarks in terms of serious harm.

Suppose we accept that the complete account of the permissibility of harming can given solely in first order terms, and that this account is built around the pma. B1 shows us that an agent can bring about a serious intended harm—broken ribs, for example—and yet be fully responsive to the essence of serious harm, at least as its essence is understood by this first order account of the permissibility of harming. A defender of the pde might claim that this leaves something out: it leaves out the fact that intending serious harm is a distinctive wrongmaking feature of action. But to someone who thinks that a first order account exhausts the truth about permissibility and impermissibility, this just begs the question. And it is only if the question is begged in this way that the claim that intending serious harm is swimming head-on against the normative current has any support. Someone who thinks that the content of the correct account of the permissibility of harming is entirely first order and is built around the pma will think that the phrase "swimming head-on against the normative current" nicely describes the Iago's of the world ("Evil be thou my good"), but that it is simply a false description of the agent in B2. So nothing in the passage gives anyone who started out by accepting the first order/second order picture any reason to question this picture, and hence any reason to accept the pde.

A second difficulty is that even if the passage were defensible, it would not seem to provide any support for the pde. Suppose that we accept that from the perspective of the agent, the intention magnifies the badness of the harm, and that this provides an acute sense of moral dislocation. (For all I know, this is an accurate empirical description of the moral psychology of decent human moral agents.) Surely if one is going to accord this any significance,

the natural reaction would be to see it as relaxing the demands morality puts upon the agent. It is not plausible to claim that intention magnifies the badness of the harm from the point of view of the victim: which would you rather suffer, a smaller intended harm or larger foreseen harm? Hence it is only from the agent's point of view that intention could plausibly magnify, and the only plausible way of construing that is that it is worse for a normal moral agent—more unpleasant, perhaps—to bring about intended harm than to bring about merely foreseen harm: the phenomenology in Nagel's example of twisting a child's arm to bring about a good certainly supports this proposal. But if this is the significance of the way intention magnifies, then if it is reflected at all in the account of the permissibility of harming, the upshot would be that in some cases, bringing about merely foreseen harm is morally obligatory, while bringing about intended harm in otherwise similar circumstances is permissible but not obligatory. But this makes the significance of intention a source of options, not constraints, contrary to the pde.

In discussing the example which drives Nagel's account, Korsgaard says: "Nagel is mistaken when he emphasizes that you are aiming at your victim's *evil*" and then immediately adds the claim: "The problem is that you are treating your victim as a mere means" (Korsgaard 1993, 47). And she goes on to develop an account which, if the added claim is correct, seems to avoid the difficulty that Nagel's own claims, if correct, would motivate options rather than constraints. Nagel's example involves you twisting a small child's arm in order to get the child's terrified grandmother, who has locked herself in the bathroom, to tell you where the keys are to her car so you can get some badly injured accident victims to the hospital. Is this using the child, or treating it as a mere means? The account of what it is to use someone I sketched earlier is based on Korsgaard's own work (Korsgaard 1986), and she repeats it in summary form in her discussion of Nagel: "[Y]ou treat someone as a mere means whenever you treat him in a way to which he could not possibly consent" (Korsgaard 1993, 45). Now I think it at least arguable that a small child could not possibly consent to what you are trying to do. You are proposing to bring about a highly structured complex sequence of events which I doubt a small child could understand at the best of times, and certainly not when you are towering over it, scaring it out of its wits; and even if it could understand, it is hard to claim that it could genuinely consent in such terrifying conditions. So it is at least arguable that you are using the child as a means (although whether the Kantian injunction not to treat people merely as means applies to small children, who are far from fully rational, is another matter). But this is just an artifact of the example: change the child into an adult who understands the situation and realizes he can only scream effectively if he has his arm twisted and consent surely is possible and you would not be using him as a means in the relevant sense even if you did not in fact obtain his consent. And if that is somehow problematic, recall that by the time we get to examples as mundane as B2, it is plainly not true that the victim is being treated

merely as a means in the Kantian sense. At most, there is a certain amount of overlap between the cases the pde condemns and the cases the Kantian injunction condemns, but the pde and the Kantian injunction are conceptually quite different, and the injunction does not provide any support for the pde.[13]

The Mismatch Principle

If I am right in thinking that the arguments that have been offered in favor of the pde do not provide anyone who accepts the first order/second order picture to start with any reasons to accept the pde, we may seem to have arrived at the point where we have to make a decision about whether the weight we attach to the first order/second order picture outweighs the weight we attach to the judgments about cases that have been offered in support of the pde. Different theorists will go different ways, but there is tension whichever way we go. It would be better to have a principle which respected the first order/second order picture and which made sense of the judgments at least as well as the pde.

The principle I want to develop is motivated by a feature of B1 and B2 since these are the cases around which the defenses of the pde falter. The agents in both B1 and B2 accept a first order complete account of the permissibility of harming and allowing harm which includes the pma. Their proposed end is saving the child, but achieving this end involves a wrongmaking feature of action, harming the bystander, and they ask whether or not the benefit to the child outweighs this harm. They conclude that it does, and pursue their end only on those grounds. There are no possible circumstances in which their reasons for action commit them to harming the bystander when that would be impermissible on the first order account of harming, as it would be if throwing the life-preserver would make the bystander drown. I suggest that because of this there is no second order moral failure in either of these cases. My main proposal will simply be the flipside of this.

To start, I propose that the content of the correct or true account of permissible and impermissible behavior is purely first order (and more or less matches the structure of commonsense morality). For short, I will simply call this account 'morality'. Say that an agent fails to act from morality in acting for reasons R if and only if there are possible circumstances in which R would commit the agent to acting impermissibly. I then propose that there is an important second order moral failure when there is a certain sort of mismatch between the agent's reasons for action and the content of morality. More precisely, the mismatch principle (mmp) says that there is a genuine and distinctive kind of second order moral failure when and only when an agent fails to act from morality. And I also propose that the pde cases in which we judge that something is morally amiss are those and only those that are condemned by the mmp, and that this proposal sorts these cases better than the pde does.

Before looking at the difficult pde cases, let's look at a simple example. In B3, the agent's sole reason for throwing the life-preserver at the bystander is to hurt the bystander. As it happens, the agent knows that hitting the bystander will result in the life-preserver getting to the child, thereby saving it. And as it also happens, the agent knows that the harm to the bystander is not so great that it is impermissible to hit the bystander. But the agent could not give a damn about all this, and throws the life-preserver simply to hurt the bystander. If the morality of harming is purely first order, this is permissible. But because the agent would have hit the bystander regardless of whether the child was there, there are possible circumstances in which the agent's reasons for action would commit her to behaving impermissibly. So by contrast with B1 and B2, the agent in B3 fails to act from morality, so she is condemned by the mmp.

To defend the mmp we will have to do two things. We will eventually have to show that failing to act from morality deserves to be regarded as a genuine and distinctive kind of second order moral failure. But first, we have to show that the mmp sorts the pde cases properly.

Cases

Since the mmp is being offered as an alternative to the pde, the most severe test of its ability to handle the cases is likely to be in the cases which seem to provide the greatest support for the pde. These are war cases and dilemmatic saving cases.[14] The most famous case in the pde literature is this.

> *Bombing* We are fighting a just war. In both Terror Bomber and Strategic Bomber, a bomber flattens the enemy munitions factory, destroying it. This weakens the enemy's war effort. The explosion also kills many innocent civilians nearby. The bombers knew all this was going to happen when they bombed the factory. In Strategic Bomber, destroying the munitions factory and thereby weakening the enemy's war effort was the bomber's reason for bombing. In Terror Bomber, killing the civilians and thereby demoralizing the enemy to weaken the enemy's war effort was the bomber's reason for bombing.

Very many of us judge that there is something seriously morally amiss in Terror Bomber but not so in Strategic Bomber. If this judgment is correct, it can only be because of the differences in the agents' reasons for action. Friends of the pde take this as support for the pde. What does the mmp say? We need to know more about the bombers' reasons for action. Here are two natural stories.

Strategic Bomber: "We need to weaken the enemy's war effort. One way of doing this is by destroying the enemy's munitions factories. So let's try to do that. Aha, there's a munitions factory within bombing range at X. But this will involve killing innocent civilians. Is this morally acceptable? Well, there

are no other comparable munitions targets, and the war is sufficiently horrible and our cause sufficiently just for it to be acceptable."

Terror Bomber: "We need to weaken the enemy's war effort. One way of doing this is by demoralizing the enemy, and one way of doing that is by killing innocent civilians. Is this morally acceptable? Well, the war is sufficiently horrible and our cause sufficiently just for it to be acceptable to kill the innocent civilians. So let's try to do that. Aha, there's a suitable population center within bombing range at X. Bombs away!"

Lots of variations are possible, but these stories are about as simple as they can be while attributing to the bombers reasons for action which fit with only the terror bomber intending the death of the civilians at the same time as telling a plausible story about what the bombers see as justifying killing the civilians. And it is what they see as providing this justification which forms the key difference between their reasons for action. Each of the bombers deliberates about what, if anything, overrides the wrongmaking feature of killing innocent civilians, then acts on this. The strategic bomber's reasons for action commit him to killing innocent civilians only in counterfactual circumstances quite close to the actual world: circumstances in which something similar to a munitions factory is destroyed and in which there is no more suitable target (more destruction to munitions or fewer civilians killed). The terror bomber's reasons for action commit him to killing innocent civilians in a much wider range of counterfactual circumstances. In particular, the terror bomber's reasons for action commit him to killing innocent civilians who are far removed from anything of strategic value to the enemy in circumstances in which that is the most efficient way of demoralizing the enemy by killing innocent civilians.

For the mmp to condemn only the terror bomber, we need to accept two claims. (1) In a just war, it is permissible to kill innocent enemy civilians if this destroys something of significant strategic value to the enemy and there is no more suitable alternative (for example: the same strategic value and fewer innocent civilians killed). (2) Even in a just war, it is impermissible to kill innocent enemy civilians if this does not destroy anything of significant strategic value to the enemy.[15]

In some form or other, both of these claims seem very plausible. Claim (1) is surely plausibly based on some kind of appeal to self-defense on a large scale. Claim (2) is surely plausibly based on the idea that there are limits to the amount of force it is acceptable to use even in self-defense. What is going on here is the view that some sorts of wars are far worse in terms of costs to human life and civilization than others, that innocent civilian lives count for a lot even on the other side, and that engaging in terror bombing is likely to provoke tit-for-tat retaliatory responses which greatly escalate the cost of the war. Furthermore, the effects of bombing isolated civilian populations are highly uncertain: it may make the enemy band together more, be less likely to sue for peace, be a more uncertain ally should we reach peace, and so on. Some-

thing along the lines of claims (1) and (2) is surely correct, but if we accept this, we have enough to claim that even though the terror bomber behaved permissibly, there are possible circumstances in which his reasons for action would commit him to behaving impermissibly. Hence the terror bomber failed to act from morality and is hence condemned by the mmp, whereas the strategic bomber did not so fail, and is not condemned. The mmp sorts the cases the way we need.

Consider now

> *Conflict* Six people lie before a threat of death, one in one location and five in another location. The doctor can treat the one by giving her a drug or the five by giving them a possibly different drug, but because of their locations the doctor cannot treat all six. Those who are not treated will die shortly, whereas those who are treated will have normal lives. There are no relevant differences between any of the six: none is any more responsible for being before a threat of death than any of the others; they are all the same age; if saved, they will have equally good outcomes; and so on. The conflict doctor decides to treat the five on the grounds that there are no relevant differences between the one and the five and that, in such cases, morality requires saving the greater number.

> *Organs* Six people again lie before a threat of death. The doctor can treat the one by giving her a drug, but the five are facing organ failure and the only way the doctor can treat them is by letting the one die (who happens to have the only compatible organs) and using her organs to save the five. As in Conflict, none is any more responsible for being before a threat of death than any of the others; they are all the same age; if saved, they will have equally good outcomes; and so on. The organs doctor lets the one die in order to save the five on the grounds that in such cases, morality requires saving the greater number.

Foot uses a pair of cases much like this to motivate the pde.[16] I think most of us judge that there is something seriously morally amiss about Organs which is not present in Conflict. Moreover, the pde seems to provide a natural explanation: the organs doctor intends the death of the one, whereas the conflict doctor merely foresees the death of the one. In my view, however, it is far from clear that the organs doctor really does intend the death of the one, so it is far from clear that the pde really does a good job in capturing our judgments about this pair of cases. But I will ignore this and ask instead whether the mmp can capture our judgments.

One interesting feature of Conflict and Organs is that unlike Strategic Bomber and Terror Bomber, Conflict and Organs are not quite perfectly symmetrical in features other than the agents' reasons for action. This should make us suspicious: it may well be that the asymmetry is the right place to look for an explanation of our differing judgments. The difference, of course, is that in

Conflict, the five need a drug of some sort or other, whereas in Organs, the five need the organs of the one. But there is an obvious asymmetry: the organs are the property of the one but the drug the five need is not. Let us see if we can develop this idea.

Perhaps morality requires the Conflict doctor to give the one a chance of being saved. But for simplicity I will assume that the Conflict doctor's reasons for action are straightforwardly correct.[17] Given the Conflict doctor's reasons for action, there is therefore no moral failure according to the mmp. What about Organs?

First consider Organs 2. Here the one is not facing a threat of death, but the five are. Do the five have any claim on the one that she provide them with her organs? I take it that they do not. Given the pma, the constraints on allowing harm are weak, and the one is certainly not, other things being equal, obliged to save the five given that the cost to her, death, would be so high. Furthermore, commonsense morality says that the constraints on harming are strong, and the doctor is certainly not, other things being equal, permitted to kill the one to save the five. Thus it is very natural to speak of the one's organs as her property.[18] Of course, what lies behind pma is controversial, but it is not controversial that this is the structure of commonsense morality.

Now consider Organs 3. This time, the one is facing a threat of death, as it happens because her organs have fallen out. The doctor can either save her by giving her a drug which will suck her organs back in, or he can take her organs and give them to the five. I take it that commonsense morality regards it as obviously impermissible for the doctor to give the one's organs to the five because they are *her* organs. But Organs 3 is just a special case of Organs. Thus the natural explanation of what is morally amiss in Organs is that it is morally equivalent to Conflict if and only if the one's organs are like the drugs in Conflict, a resource to which, to put it very roughly, equal need creates an equal claim.[19] But the organs are not like this: the five have no claim to them but the one has a very strong claim; they're hers. Given the situation the five are in, they have a claim to be saved by the doctor only if they have a claim to the one's organs. Since they do not have this claim, only the one has a claim to be saved, so morality requires the doctor to save the one. Hence the conflict doctor acted impermissibly. Moreover, the conflict doctor's reasons for action committed her to acting impermissibly, hence she failed to act from morality, and so is condemned by the mmp.

Note also that it looks very implausible to claim that the organs doctor intended the death of the one in Organs 3. It therefore looks implausible to claim that the pde entails that there is anything morally amiss about Organs 3. But since Organs 3 is very similar to (in fact, a special case of) Organs and we judge that there is something seriously morally amiss about Organs 3, it seems natural to suppose that the explanations of what is amiss in both Organs 3 and Organs will be very similar. The pde's claim to be providing a good account of Organs is thereby weakened.

In short, if we accept that the way commonsense morality treats harming and allowing harm is correct, we can explain our judgments about Conflict and Organs just as well, and probably better, via the mmp than via the pde.

Consider now

> *Learning* One person has one life-threatening disease while five people have a different life-threatening disease. The one and the five are the same in all relevant respects, including the fact that they will die unless their diseases are treated. The doctor is able to treat the one via procedure A, but the five need to be treated via procedure B, and it is not known how to do this. The only way the doctor can discover how to save the five via procedure B is by letting the one die and learning from her case. The doctor decides to let the one die on the grounds that in such cases, morality requires saving the greater number.

This is a modification of a pair of cases offered by Quinn.[20] Most people find something morally amiss in this case, and again, it appears to provide support for the pde. Moreover, it appears to be more plausible (though in my view, not obviously correct) to claim that the doctor intended the death of the one in this case. But my sense is that the structure of our explanation of Organs carries over to this case. The five can be saved only if the one provides a certain resource, in this case information. But if we build Learning 2 out of Learning in the same way we built Organs 2 out of Organs, it is clear that commonsense morality says that the five do not have a claim against the one that she provide the information, whereas the one does have a claim against the doctor that she save him. The explanation then runs in parallel, and we can conclude that the learning doctor behaved impermissibly. Moreover, the learning doctor's reasons for action committed her to acting impermissibly, hence she failed to act from morality, and so is condemned by the mmp.

Overall, Conflict, Organs and Learning illustrate a somewhat trivial application of the mmp in which the possible circumstances in which the agent's reasons for action commit her to acting impermissibly include the actual circumstances. In such cases there are two moral failures on the view being developed here, first order and second order. But as Bombing illustrates, there can be a second order moral failure without a first order moral failure.

Comparison

Let's now compare the pde and the mmp. The cases we have looked at are the cases the defenders of the pde have seen as providing the best support for the pde. It is therefore likely that they are cases that provide the best support for the pde, and therefore the toughest cases for an alternative to the pde to handle. As I indicated, I am inclined to think that the pde has some difficulty with condemning the saving cases where and only where we judge that there

is a moral failure because of the implausibility of the claim that the doctor in Organs 3 intended the death of the one, but let us give the pde the benefit of the doubt. Even so, the mmp sorts these cases at least as well as the pde.

In fact, one would expect the pde and the mmp to deliver similar judgments in a wide range of cases. In accepting the pma, commonsense morality says that there are strong constraints on harming. To a first approximation, this means that for it to be permissible to harm someone, the benefits have to greatly exceed the magnitude of the harm, and if the harm is serious enough, usually no benefits would be great enough. We might give this a rough statistical summary by saying that harming is usually impermissible. Now one difference between bringing about harm which one merely foresees and intending the harm is that there is typically a much wider range of counterfactual circumstances in the intending case in which one is committed to harming in virtue of one's reasons for action, and only one of them has to be impermissible for this to count as failing to act from morality. We might give this a rough statistical summary by saying that if harming is usually impermissible, intending harm almost always involves a failure to act from morality. Hence although the pde and the mmp are conceptually very different, in practice almost all the cases condemned by the pde are condemned by the mmp.

However, there are three important classes of cases over which the pde and mmp diverge. The first is made up of cases like B2 in which an agent intends harm—hence is committed to harming a wide range of possible circumstances—without being committed to behaving impermissibly (according to a first order account). The pde says that cases like this involve a distinctive kind of wrongmaking feature of action. The mmp denies that there is any kind of moral failure in action in these cases. For example, the mmp says that there is no moral difference between B1 and B2, but the pde says that B2 involves a distinctive wrongmaking feature of action. I believe this disagreement between the mmp and pde supports the mmp in two ways. First, along with those I have consulted, I judge that there is no moral difference between B1 and B2. And even some writers who seem to have some sympathy for the pde accept that there are otherwise identical intended/foreseen pairs of cases between which there is no moral difference.[21] Second, recall that it is when applied to cases in this class that the well known defenses that have been offered for the pde fail. It is surely no coincidence that the pde does not make sense of our judgments about cases like B2 *and* that the pde's defenses fall apart when applied to such cases. This suggests that cases like B2 constitute a morally important subset of intended harm cases, and while the mmp explains this, the pde does not. So overall, cases like B2 support the mmp.

It would not be too unfair to summarize the flavor of the discussion of cases so far as "what the pde can do, the mmp can do better". But this disguises the difference between the pde and the mmp. The mmp in no way accepts that intending harm constitutes any kind of moral failure in action. At most, the mmp condemns a fairly wide range of cases involving intended harm. In

fact, the mmp entails that there can be otherwise identical intended/foreseen cases in which the foreseen case involves a distinctive kind of moral failure in action but the intended case does not. Consider B4, a variant of B1. Recall that in B1 the agent hits the bystander on her backswing, but she does this only because, in part, she judges that the harm to the bystander is not so great that it is impermissible to hit the bystander if this will save the child. In B4, the agent attaches no negative moral weight at all to hitting the bystander on her backswing; it's her child, and even if it is facing only a small risk of drowning, she couldn't care less about the bystander drowning. After all, he is not even white. As it happens in B4, it was clear that the harm to the bystander would be minimal and that the child was drowning, so the agent's behavior was permissible. But it is clear that the agent's reasons for action were seriously objectionable. The mmp agrees as the agent is committed to harming the bystander in possible worlds in which the harm to the bystander would be very serious while the risk of her child drowning is minor. But the pde has no esources for saying that there is any kind of moral failure in action in B4, so it does not capture our judgments about cases like B4. Thus the second class of cases over which the pde and the mmp diverge is made up of merely foreseen cases like B4, and again, only the mmp captures our judgments about these cases.[22]

When we combine the first and the second classes of cases, we get the interesting result that there can be otherwise equivalent intended/foreseen pairs like B2 and B4 in which only the foreseen case involves a distinctive kind of moral failure in action. This illustrates how intending harm constitutes no kind of distinctive moral failure in action at all according to the mmp, and how this matches our judgments about cases.

"But isn't there is still something morally objectionable about intending harm?", some will ask. I ask where the evidence is. Defenders of the pde have offered us only a remarkably narrow range of intended/foreseen pairs of cases—I am aware of only about four or five—in which, not surprisingly, the pde looks at its best. But as far as I am aware there has been no attempt to describe the many forms intended/foreseen pairs can take and to catalogue the widely shared reactions (if such there be) to these pairs. You would hardly be convinced of the virtues of a new drug by being told by the manufacturer of four or five cases in which the patients recovered, and I do not see how the evidence base for the pde is much stronger. In fact, we have now seen three sorts of intended/foreseen pairs which differ only in terms of the agents' reasons for action. In the first sort we judge that there is a moral failure in the intended cases only, in the second sort we judge that there is no moral failure in either, and in the third sort we judge that there is a moral failure only in the foreseen case. The pde and mmp agree about the first sort, the pde gets it wrong about the second while the mmp gets it right, and the pde is silent about the third while the mmp gets it right. Since our judgments about intended/foreseen pairs vary in ways that are tracked by the mmp but missed by the pde, I do not see where the case based evidence for the pde is supposed to lie.

The third class of cases over which the pde and mmp diverge involves cases like Kant's shopkeeper. Recall that this shopkeeper notices the chance to make a quick profit by giving an inexperienced customer the wrong change. But he gives the right change only for the reason that he will make more profit in the long run that way. If we consider a possible world where there will be no more interaction between the inexperienced customer and both the shopkeeper and other customers and people who influence other customers and so on, it is clear that the shopkeeper's reasons for action commit him to giving the wrong change in this world, hence to acting impermissibly. So in the actual world, he failed to act from morality, hence the mmp condemns him even though he acted permissibly. We noted earlier that it is possible to broaden the scope of the pde so that it condemns not just intended harm, but intended promise breaking and so on, so that it could condemn intended wrong change giving. But this is not much help since Kant's shopkeeper does not give the wrong change. Most people (Kant included) judge that Kant's shopkeeper involves some kind of moral failure in virtue of the shopkeeper's reasons for action, but the pde will not pick this up.

In summary, it is very hard to deny that the mmp is much better supported by cases than the pde. There is, however, a possible response on behalf of the pde. One problem for the pde is that it gets it wrong about intended/foreseen pairs where we see no moral difference. I do not see how the pde can overcome this difficulty, but let us ignore this. The other problem is that the pde does not condemn cases it should condemn, cases like B4 and Kant's shopkeeper. But a defender of the pde might claim that the pde does not have to claim to be the only important principle about moral failures in virtue of agents reasons for action. Perhaps there is another principle which condemns some sorts of reasons for action, picking up the slack left by the pde. I do not see how anything but an unshakeable commitment to the pde would persuade one that there is such a principle in the absence of finding it, but let us grant that something could be found. But the resulting position would suffer from disjunctivitis. The phenomenon we are trying to explain, moral failures in virtue of reasons for action, cries out for a unified explanation. The mmp provides a natural, simple, unified explanation, but the imagined pde plus something else position provides only a disjunctive explanation. So even if the imagined pde plus something else explanation could be found, the mmp would still be better supported by cases and a deeper principle.

This shows, I believe, that the mmp is far more strongly supported by judgments about cases than the pde. But the pde also faces theoretical difficulties. (1) Recent defenders of the pde have seen it as more plausible in a nonabsolute form (it is very hard to believe that it is impermissible for the agent to save the child in B2) but then we are owed an account of how the wrongmaking feature of bringing about intended harm interacts with other right and wrongmaking features of action. But the mmp only needs a first order account of right and wrongmaking features of action. The pde needs that as well, but

it then faces this additional interaction problem. Without having a theoretical rationale for the pde and hence without knowing what is so morally problematic about intended harm, it is very difficult to know how to attack this problem.[23] So the pde faces one more difficulty than the mmp. (2) Unlike the pde, the mmp respects the natural first order/ second order picture. Even defenders of the pde agree that there is an initial plausibility to this picture. Since there is not, as far as I can see, *any* reason, theoretical or case-based, to prefer the pde to the mmp, it has to count as a cost of the pde that it does not respect this picture. (3) The existing attempts to provide a rationale for the pde do not succeed. Moreover, given that many leading philosophers have searched for a defense, this is surely good evidence for the claim that there is no rationale. We can therefore finish the case for the mmp by providing it with a rationale.

Rationale for the MMP

There are really two tasks here. One is to provide a complete defense of the mmp as properly characterizing a genuine second order moral failure. The other is to give reasons to accept that such a defense could be given. Since the main goal of this article is to show that we should accept the mmp in place of the pde, and since the pde has not been given a successful theoretical defense, it will be enough to pursue the second task. This is fortunate since, as I will indicate at the end, pursuing the first task goes far beyond anything that could be done here.

The claims about the structure of commonsense morality that we have been basing our discussion of the pde and the mmp on, such as the pma, are unlikely to be defensible on a broadly utilitarian theory. I will therefore assume that we are working within some form of contractualism of a broadly Kantian kind.[24] I take it that on this kind of view, what makes a morality true or correct is that it is the set of propositions about permissibility and impermissibility of behavior which, if generally followed, would result in a roughly equal distribution of various goods needed for effective rational agency, such as a variety of freedoms and all purpose resources.[25] Someone who behaves impermissibly, on this account of morality, against the background of a community of agents generally disposed to follow that morality is almost inevitably doing so to promote her own interests or ends, and can therefore be seen as taking a kind of unfair advantage of the benefits of moral cooperation. But someone who fails to act from morality even though she is behaving permissibly is thereby committing herself to taking that kind of unfair advantage if she can; given the content of her reasons for action, it is just a matter of luck that she doesn't. If others become aware of this—as we become aware when we are presented with their reasons for action—resentment is a natural attitude to take to her, particularly on the part of her would-have-been victims or their spokes-

persons, and particularly if those would-have-been victims were disposed to keep to their role in moral cooperation.

Similarly, shame is a natural attitude for a normal moral agent to take who on a particular occasion fails to act from morality. The full story of normal moral motivation on this kind of contractualist picture is no doubt very complex, but it must have to do with seeing others as having, very roughly, an equal claim on the benefits of cooperation purely in virtue of their capacity to cooperate (and perhaps, in addition, their being disposed so to cooperate). Normal moral agents will have internalized and to some degree identified with this motive.[26] But there is an ill-fit between having this motive and failing to act from morality, and it is therefore natural for a normal moral agent who reflects on having failed to act from morality to feel shame.

In short, I share the view that the best prospects for defending the structure of commonsense morality that the mmp (and pde) rest on lie in broadly Kantian forms of contractualism. But if we accept this view, then it is very natural to associate failing to act from morality with two of the attitudes that are central to second order morality, resentment and shame. This strongly suggests that it is possible to provide a good defense of the claim that the mmp characterizes a genuine and important kind of second order moral failure, which is what we needed.

At this point, a full defense of the mmp would have to face a huge number of questions. What is the detail of the account of how an agent's reasons for action commit her to behaving in different circumstances? How do reasons for action need to be understood to get this account to work? How does the account operate in cases of what are said to involve overdetermination of reasons for action? How does failing to act from morality connect with the Kantian notion of failing to act from the motive of duty? How does the mmp fit into second order morality? Is failing to act from morality the central second order moral failure or is it more part of a network? How does it connect with the others? It would plainly be vain to address these topics here. But since one test of a good theory is its ability to generate interesting links with other topics, the fact that the mmp raises these unanswered questions may not count against it.

Notes

*An early version of this paper was delivered at the University of Melbourne, and I am grateful for comments. Still earlier I benefited from conversations with Richard Holton, Frances Kamm, Rae Langton, Derek Parfit, Philip Pettit, Julian Savulescu, Peter Singer and Michael Smith. Arthur Kuflik read an almost final version along with a draft of another paper on related material and I benefited greatly from the long discussion that followed. I received helpful comments from anonymous referees, with one referee's comments forcing major revisions in the analysis of cases.

[1] An uninteresting way: a reason for action detector will kill ten children if you go into the kitchen in the next five minutes to get coffee.

[2] I borrow the terminology 'first order morality' and 'second order morality' from (Bennett 1995, 46). Bennett attributes it to Alan Donegan.

[3] Such a weakening is endorsed in (Quinn 1991), (Nagel 1986, 185), (Kamm 1992, 376) and (Kamm 1991).

[4] The eventual need for this kind of extension is suggested in (Nagel 1986, 185).

[5] The extension to allowing harm is implicit in (Foot 1985), (Quinn 1989) and (Nagel 1986, 180).

[6] I should qualify this paragraph. The three major attempts to defend the pde are (Foot 1985), (Quinn 1989) and (Nagel 1986). Foot and Quinn accept the need for conjoining a defense of the pde with an independent defense of the pma to get a defense of commonsense morality. Nagel tentatively suggests that the pde will do most of the work by itself, but for the reasons given in the text, I do not find this plausible.

[7] (Quinn 1989) explicitly endorses this position. The magnification metaphor is also present in (Nagel, 1986).

[8] The main assumption I will rely upon is that the pma can be defended within the framework of the first order/second order picture. I argue for this assumption in (McCarthy 2000).

[9] For extended discussions of the pde on this topic, see (Bratman 1987) and (Bennett 1995).

[10] A referee asked for remarks on the place of an agent's beliefs in the first order/second order picture. The worry expressed was that beliefs seem relevant to questions about permissibility in two ways. (1) It seems a mistake to say that a lunatic behaves permissibly or impermissibly, and part of what seems to be going on here is that lunatics lack appropriate belief forming mechanisms. (2) At least some people regard it as a mistake to say that an agent behaves impermissibly when as a result of excusable ignorance she brings about some awful outcome (she turns on her gas stove and because of some bizarre causal chain triggers a fatal explosion in her neighbor's apartment). I accept versions of (1) and (2), but I claim that these are consistent with the first order/second order picture. The first order/second order picture can accommodate (1) by claiming that having sufficiently sophisticated and reliable belief forming mechanisms is a necessary condition for a creature to be the kind of creature whose behavior can properly be said to be permissible or impermissible. But this does not entail that what it is permissible for an agent whose behavior can be so evaluated to do is a function in part of the agent's beliefs about the consequences of her actions. The first order/second order picture can accommodate (2) by saying that insofar as the consequences of an agent's behavior are relevant to the permissibility of her behavior, it is the expected consequences that are relevant, where the notion of probability used to define 'expected' is suitably objective. This makes the evidence about what the consequences of her actions would be that is available (and not available) to the agent and the objective probabilities that evidence supports relevant to the permissibility of the agent's behavior, but it does not make her beliefs about the consequences relevant. In the example given, there was no evidence to suggest that turning on the gas would result in an explosion, hence the agent did not behave impermissibly. I argue for this position in (McCarthy 1998). Of course, it is possible to accept the first order/second order picture and deny (2), as some do. My only point here is that accepting the first order/second order picture is consistent with (2).

[11] A similar point is made in (Bennett 1995, 221–222) and (Quinn 1989, 347).

[12] See (Quinn 1989). For discussion, see (Kamm 1992), (Fischer, Ravizza, and Copp 1993) and (McMahan 1994).

[13] In fairness, Korsgaard does not say anything about the pde. Nagel's discussion attempts to do two things, provide a defense of the pde and show that the pde provides deontology, which I will take at least approximately to be the pma. As I indicated earlier, I am skeptical about whether a defense of the pde would generate a defense of the pma, at least in the version accepted by commonsense morality. But Korsgaard is only interested in reconstructing Nagel's arguments for deontology and there is no real evidence that she believes that this reconstruction will also provide a defense of the pde.

[14] Some have discussed applications of the pde to abortion and euthanasia. In my view, these are not good cases to try to motivate the pde with. In standard euthanasia cases death is good for

the person involved, hence not a harm, hence the pde does not even get a grip on such cases. Or if it is somehow made to get a grip, it does not look very plausible. Likewise, in standard abortion cases it is at least arguable that death is not a harm to the fetus, and there is too much else going on besides for these cases to be helpful in a discussion of the pde. I will therefore ignore these cases. I also ignore the trolley problem. Some have tried to bring the pde to bear on this problem, but I side with those who see the solution to the trolley problem as lying in the (first order) account of harming and allowing harm. For attempts at such solutions, see (Thomson 1990, Ch.7) and (Kamm 1996, Chs. 6 and 7). Moreover, there are strong reasons to believe that the pde cannot be all of the truth about the trolley problem. See e.g. (Kamm 1992, 380–81).

[15] I ignore issues about whether it is impermissible only conditional upon the enemy not doing the same to us.

[16] See (Foot 1985, 23–25). My earlier analysis of these cases collapsed under forceful criticisms from an anonymous referee for which I am very grateful.

[17] If you think that the Conflict doctor was morally required to give the one some chance, then assume that both she and the Organs doctor give the one that chance on the grounds that in such cases, morality requires giving the one such a chance. The arguments that follow will go through unchanged.

[18] For defense of the idea that we have property rights in our bodies, see (Thomson 1990, 225–26).

[19] This slogan is, I believe, consistent with the idea that claims can aggregate in a morally significant way, making it the case that the conflict doctor's reasoning is correct.

[20] See (Quinn 1989, 177). The modification follows a suggestion of an anonymous reviewer who pointed out, correctly I believe, that my treatment of cases more similar to Quinn's originals did not adequately explain what was amiss about this case. Although the reviewer did not say this explicitly, the new version appears to provide more powerful support for the pde than the earlier version.

[21] See, for example, the discussion of the Prevented Return Case, a variation of the Trolley Problem, in (Kamm 1992, 380–381). Kamm discusses this case further in (Kamm 1996).

[22] I am grateful to an anonymous referee for pointing out the need to discuss these kinds of cases.

[23] This problem is noted in (Kamm 1991) and (Kamm 1992).

[24] Of course it is controversial whether the pma, for example, is defensible even within such forms of contractualism. I try to show it is in (McCarthy 2000).

[25] For the Kantian pedigree see (Herman 1985) and (Rawls 1980).

[26] Compare the picture of Kantian moral psychology developed in (Herman 1991).

References

Bennett, Jonathan. (1995) *The Act Itself*, [Oxford: Clarendon Press].

Bratman, Michael. (1987) *Intention, Plans, and Practical Reason*, [Cambridge: Harvard University Press].

Fischer, John Martin, Ravizza, Mark and Copp, David. (1993) "Quinn on Double Effect: The Problem of 'Closeness'," *Ethics* 103: 707–725.

Foot, Philippa. (1985) "Morality, Action, and Outcome" in (Honderich 1985).

Herman, Barbara. (1991) "Agency, Attachment, and Difference," *Ethics* 101: 775–797.

Herman, Barbara. (1985) "The Practice of Moral Judgment," *Journal of Philosophy* 82: 414–436.

Honderich, Ted ed. (1985). *Morality and Objectivity*, [Routledge & Kegan Paul].

Kamm, Frances. (1996) *Morality, Mortality* Vol. II [New York: Oxford University Press].

Kamm, Frances. (1992) "Non-consequentialism, the Person as End-in-Itself, and the Significance of Status," *Philosophy and Public Affairs* 21: 354–389.

Kamm, Frances. (1991) "The Doctrine of Double Effect: Theoretical and Practical Issues," *Journal of Medicine and Philosophy* 16: 571–85.

Korsgaard, Christine. (1993) "The Reasons We Can Share: An Attack on the Distinction Between Agent-Relative and Agent-Neutral Values," *Social Philosophy and Policy* 10: 24–51.
Korsgaard, Christine. (1986) "The Right to Lie: Kant on Dealing with Evil," *Philosophy and Public Affairs* 15: 183–202.
McCarthy, David. (2000) "Harming and Allowing Harm," *Ethics* 110: 749–779.
McCarthy, David. (1998) "Actions, Beliefs, and Consequences," *Philosophical Studies* 90: 57–77.
McMahan, Jeff. (1994) "Revising the Doctrine of Double Effect," *Journal of Applied Philosophy* 11: 201–212.
Nagel, Thomas. (1986) *The View from Nowhere*, [New York: Oxford University Press].
Quinn, Warren. (1991) "Reply to Boyle's 'Who is Entitled to Double Effect?'," *Journal of Medicine and Philosophy* 16: 511–514.
Quinn, Warren. (1989) "Actions, Intentions, and Consequences: The Doctrine of Double Effect," *Philosophy and Public Affairs* 18: 334–351.
Rawls, John. (1980) "Kantian Constructivism in Moral Theory," *Journal of Philosophy* 77: 515–572.
Thomson, Judith. (1990) *The Realm of Rights*, [Cambridge: Harvard University Press].