

## Moral Inertia<sup>\*</sup>

### **1. Introduction: Deflection**

The focus of this paper will be scenarios of *deflection*, i.e. scenarios where an agent interferes with an ongoing process (a process already “in motion”) by making it depart from its preset path and take a different path. Cases with this structure have shown to be of significant interest to both metaphysics and ethics. In metaphysics, deflection scenarios have been a recent focus of investigation by philosophers working in the metaphysics of causation. For some deflection scenarios—temporary deflections of processes, where the processes then resume their initial path—suggest that it is possible to affect the causal route to an outcome without causing the outcome (say, I temporarily divert a missile, which then resumes its original path and reaches its destination all the same; it seems that the deflection doesn’t cause the outcome). The challenge, then, is to say what distinguishes those arguably non-causal deflection structures from other structures that are arguably causal (in particular, “preemption” scenarios, such as the launching of a new threat that goes to completion before an old threat does).<sup>1</sup> In ethics, deflection cases are the source of intriguing “trolley puzzles,” where the challenge is to explain the moral difference that we see between two different kinds of scenario, one of which is a deflection scenario. For example, if a runaway train is threatening to kill five workmen on the tracks, it seems permissible to switch the train onto a side track where only one man is working. On the other hand, it doesn’t seem permissible to

---

<sup>\*</sup> I would like to thank the audiences at the 2007 Bellingham Philosophy Conference, the University of Wisconsin-Milwaukee, the University of Notre Dame, and Northern Illinois University, for their helpful comments on earlier versions of this paper. Special thanks to Juan Comesaña and to my commentators at the Bellingham conference, Anne Barhill and Peter Graham.

<sup>1</sup> See, e.g., Paul (2000), Yablo (2002), and Sartorio (2005).

throw a person in front of the train, even if it's to prevent the train from killing the five people on the tracks.<sup>2</sup>

In this paper I discuss another role played by deflection scenarios in ethics. I argue that certain deflection scenarios suggest that, according to ordinary morality, there is moral pressure to leave things “as is,” in other words, to fail to intervene. I call this phenomenon *moral inertia*.<sup>3</sup> I argue that moral inertia takes two main forms. The first form consists in strict prohibitions against interventions (discussed in section 2), and the second form consists in constraints or limitations on interventions (discussed in section 3). Both of these manifestations of moral inertia involve deflection scenarios where an ongoing process that was bound to harm or benefit some people can be diverted onto other people, thus changing who is benefited or harmed.

Moral inertia is likely to be connected to the thesis that there is a moral difference between killing and letting die, or between doing and allowing harm. I discuss the relation between moral inertia and the killing/letting die distinction in connection with each of the two manifestations of moral inertia, in sections 2 and 3.

Note that the main thesis of this paper is a claim about that to which we are *ordinarily* committed given our intuitions about certain scenarios, not a rational defense of such intuitions. However, there is a particular objection to the rational defensibility of moral inertia that has to do with the intelligibility of the concept of deflection in general. In section 4 I consider this objection and I explain how it might be possible to address it by appeal to the

---

<sup>2</sup> See Thomson (1976).

<sup>3</sup> Following Thomson's usage in her (1976) (Thomson, as we'll see, rejects the existence of moral inertia). This concept of moral inertia ought not to be confused with a different, quite common usage of the expression: the “moral blindness” to evils generated by habit.

role played by deflection in metaphysics, which was discussed at the beginning of this section.

## **2. First manifestation: Prohibitions on interventions**

### **2.1. Malm and killing versus letting die**

The first kind of pressure to leave things “as is” arises in deflection scenarios of the following kind: a process that is already in motion is about to harm or benefit one person, and you can deflect it so that another person is harmed or benefited instead. In particular, the same number of deaths would result from your intervening or not intervening, but different people would die in each case. In this important respect, these cases are different from typical “trolley scenarios,” where deflecting an ongoing threat would result in *fewer* deaths.

Heidi Malm has offered deflection cases of this kind as an argument for the moral relevance of the (ordinary) killing/letting die distinction.<sup>4</sup> Although my focus here is the distinction between intervening and failing to intervene, not the distinction between killing and letting die, the two distinctions seem to be tightly connected to each other. At least generally, we think that an agent “merely” lets someone die when there is a deadly process already in motion and the agent fails to intervene or interfere with it. By contrast, we think that an agent kills someone when he intervenes in a way that results in someone’s death, either by starting a new threat to the person or by helping to sustain an already existing threat. I won’t try to decide the issue of whether the two distinctions completely overlap, or whether

---

<sup>4</sup> Malm (1989). The thesis that she argues for is only that sometimes there is a moral difference between killing and letting die even if all other things are equal (although there are other cases where other things are equal and there is no difference between killing and letting die). As indicated, Malm’s focus is the ordinary distinction between killing and letting die; she doesn’t attempt to give a theoretical account of the distinction. Similarly, my focus is the ordinary distinction between intervening and failing to intervene.

there are cases in which they come apart. It seems to me that Malm's argument can be seen as an argument for moral inertia, even if it's also an argument for the moral significance of killing versus letting die. In fact, if, as suggested, the killing/letting die distinction rests on the distinction between intervening and failing to intervene, it seems preferable to regard the argument as an argument for moral inertia than as an argument for the moral significance of the killing/letting die distinction. (Also, see section 3 below for a possible reason to think that the thesis of moral inertia has a broader scope than the thesis about killing and letting die.) In what follows I use "killing versus letting die" and "intervening versus failing to intervene" interchangeably, unless I explicitly indicate otherwise.

To my mind, deflection scenarios provide the strongest possible case for the moral significance of the killing/letting die distinction. In what follows I make use of one of Malm's examples and I endorse Malm's reasoning about it. I believe, however, that Malm fails to recognize (or sufficiently emphasize) the important role that deflection plays in the argument. By contrast, I will argue that deflection plays a key role—actually, a *twofold* role. As we will see, it is in virtue of this double role that deflection scenarios provide the strongest possible case for the moral significance of the killing/letting die distinction.<sup>5</sup>

Consider a deflection scenario involving an evil or threat:

*Runaway Train:* A runaway train is hurtling down the tracks when it reaches a switch.

X was walking on the left-hand track when his foot got stuck, and the same happened

---

<sup>5</sup> There are other important differences between my proposal and Malm's, which will be apparent later (see, in particular, n. 18 below).

to Y on the right-hand track. The switch is originally set for the left-hand track. You could flip the switch and divert the train to the right-hand track.<sup>6</sup>

And here is a deflection case involving a good:

*Floating Drug*: Two people, X and Y, need a certain drug to survive. X, Y, and a bottle with a dose of the drug are in the water, and the bottle starts drifting towards X. You could alter its course and deflect it to Y.<sup>7</sup>

My intuition about these cases is clear: other things being equal, you ought *not* flip the switch in Runaway Train and you ought *not* deflect the drug in Floating Drug. Intuitively, we would think: you need a reason to intervene, and there isn't such a reason.<sup>8</sup> In particular, the same number of people would die if you intervened, so you wouldn't be minimizing the number of deaths by intervening—unlike, notably, in standard trolley scenarios where a threat can be deflected so that fewer people die. On the other hand, if you didn't intervene, we wouldn't blame you for not intervening (for, again, the same number of people would have died if you had intervened). Of course, it was just *bad luck* for X (in Runaway Train) that his foot got stuck on the left-hand track and Y's on the right-hand track, and not vice-versa, in other words, it was bad luck for X that he was originally threatened by the train and Y wasn't. But, then again, bad luck happens (and so does good luck). Now, if this is right, there is *moral inertia*: sometimes how things are (or were bound to be) determines what we can permissibly

---

<sup>6</sup> This case is similar to one of Malm's own examples (Malm (1989), p.238).

<sup>7</sup> This is a variation on Judith Thomson's "health-pebble" example in Thomson (1976), p.84. Thomson's example is importantly different in that the health-pebble could be deflected from five people to one person.

<sup>8</sup> Malm (1989), p. 248.

do. In particular, the preexistence of a threat—or, more generally, of a causal process of some sort—makes intervening impermissible, if other things are equal.

To bring out this intuition more clearly, think about what X could try to do or say to convince you to flip the switch in Runaway Train (in those few seconds left before the train runs him over). He could cry out for help and hope that you haven't noticed Y's presence on the other track. Or, if it is obvious that you are aware of Y's presence, he could try to persuade you that his life is more important than Y's life. But he couldn't do much more than this.<sup>9</sup> By contrast, it seems that Y has more convincing reasons to persuade you to not flip the switch. If he sees that you're about to flip the switch, the first thing that will probably cross his mind is that you haven't noticed his presence on the track and thus he will alert you to it. If he sees that there is still a chance that you might flip the switch, he'll surely complain that it's unfair for you to do so because he was never under the threat of the train, but X was (however tragic this might be for X). If all other things are equal, then, it seems that it would be unreasonable for X to expect that you'll flip the switch but it would be reasonable for Y to expect that you won't flip it.<sup>10</sup> Presumably, the same goes for Floating Drug: if other things are equal, there is no good reason to divert the drug from X to Y, so you shouldn't intervene.

Now, you might wonder whether this is too strong. What if the bystander by the switch panics and, being at a loss about what to do, and with the best intentions, flips the switch? (Perhaps, while in such panicky state, he flips the switch back and forth, and the train happens to go through the switch when it's set for the right-hand track.) Do we really want to

---

<sup>9</sup> Could he demand that you flip a coin or use some other random decision procedure to decide who dies and who lives? I consider this possibility below in the text.

<sup>10</sup> Similarly, Malm argues, Y's relatives would have legitimate grounds for complaint if you decided to flip the switch but X's relatives would not, if you decided not to do so (Malm (1989), p. 246).

say that the bystander acted wrongly?<sup>11</sup> I think that, in fact, we do. That was the wrong thing to do from an objective perspective. Of course, the agent's immoral behavior can still be excusable: perhaps the facts of the situation weren't fully clear to him at the time, perhaps everything happened so fast that he didn't have time to think it through, perhaps everyone else would have had similar trouble figuring out what to do in a similar situation. But this doesn't change the fact that his act was wrong. It seems to me that, were the facts completely apparent to us at the time, most of us would not flip the switch and would consider flipping the switch, and thus harming Y, morally unjustifiable.

A possible reaction is that this is still too strong because it rules out the possibility of flipping a coin. How about flipping a coin to decide who dies and who lives? Is this impermissible too? My intuition is the same: even this is wrong. Again, you would need a good reason to intervene (in this case, by flipping a coin), but it seems that you don't have one. For, if you have no good reason to change who lives and who dies, then, presumably, you also have no good reason to change the *chance* that the people in the situation will die. Regardless of whether the coin comes up heads or tail, one person will die. So, by flipping a coin, you are raising the chance that Y will die without thereby lowering the chance that someone or other will die.

Now, you might insist that the fact that, by flipping a coin, you're giving each person an equal chance of survival is a good enough reason to intervene in that way. I don't find this persuasive.<sup>12</sup> At any rate, it is important to realize that, even if you think that flipping a coin is permissible, the phenomenon of moral inertia still arises, although in a different form. For,

---

<sup>11</sup> Thanks to Karen Bennett for pressing this objection.

<sup>12</sup> Chance already played a role in determining X and Y's position on the tracks. So why should we leave it up to chance again? For further argumentation, see Malm (1989), section II.

even if you think that flipping a coin is permissible, you probably think that *simply flipping the switch* (i.e. flipping it not as the result of a coin-toss or a random procedure of any sort) is impermissible, but *simply failing to flip it* is permissible.<sup>13</sup> If so, there is still an asymmetry concerning certain kinds of interventions (“non-random” interventions): non-random interventions are impermissible, but non-random failures to intervene are not. In other words, if flipping a coin were permissible in situations of this type, then there would still be moral inertia, but in the form of constraints on interventions (the type of inertia discussed in the next section) rather than in the form of strict prohibitions against interventions.

Moral inertia, then, is a kind of pressure to leave things unchanged. But, when do we leave things unchanged and when do we change them? In other words, what counts as an intervention and what counts as a failure to intervene? In many cases this is intuitively clear. In other cases, however, there is no clear answer. What is our intuitive moral judgment in those cases? I turn to this in the next section.

## 2.2 Unclear cases

Consider the following variants of Runaway Train:

*The singing variant:* You are singing a tune while walking by the train tracks. When you get closer to the switch, and as the train approaches it, you see that the mechanism is equipped with a sound sensor. You realize that your singing would trigger it.

---

<sup>13</sup> People like Tooley would presumably disagree. I discuss Tooley’s view in section 2.3 below.

*The breathing variant:* You are walking, silently, by the train tracks. When you get closer to the switch, and as the train approaches it, you see that the mechanism is equipped with a breathing sensor. You realize that your breathing would trigger it. You could easily hold your breath for a few seconds until the train goes through the switch.

What should you do in these cases? Should you keep singing, or should you stop? Should you keep breathing or should you hold your breath? If you are like me, you'll probably think that in these cases there is no obvious answer—or, at least, that the answer is less obvious than in the original version of the case.

Now, the lack of clarity of our moral intuition in these cases seems to go hand in hand with a lack of clarity about when you would be killing and when you would be letting die, or about when you would be intervening with existing states of affairs and when you would be failing to intervene. Do you kill if you continue singing, or if you stop? Do you kill if you continue breathing, or if you hold your breath? Again, it's not fully clear. In fact, there seem to be two fairly persuasive lines of reasoning pointing in opposite directions. Consider the singing variant. The first line of reasoning is: the switch was originally set for the left-hand track; your singing would change the existing states of affairs in a way that results in Y's death; hence, if you keep singing, you kill Y. But the alternative line of reasoning is: you were already singing when you approached the switch; your discontinuing this behavior would change the existing states of affairs in a way that results in X's death; hence, if you stop singing, you kill X. In each case, we see a killing as involving a particular *disruption* of an ongoing process or train of events. However, there are two main ways of conceiving of the

process itself: according to one of these ways, the process includes the agent's own antecedent behavior and dispositions; according to the other way, it excludes these.

In other words, the unclarity about the distinction between intervening and failing to intervene arises from the fact that the agent (the "intervener") is, himself, part of nature. Thus, on pain of committing ourselves to an implausible anti-naturalism about human agency, we should regard agents as potential ingredients in the natural, causal processes that occur in the world. But then, given that the agent's own antecedent behavior, dispositions, etc. can in principle contribute to determining the "default" state of the world, it can be unclear what it would take for the agent to interfere with existing states of affairs, as opposed to his failing to interfere with them. As a result, our intuitive distinction between killing and letting die can fail to yield a clear verdict in some cases.<sup>14</sup>

The unclarity of our intuitions also comes in degrees. In both the singing and breathing variants, it's unclear to me when the agent would be intervening and when he would be failing to intervene. Still, I feel *more* of a temptation to regard the agent's continuing to breathe than the agent's continuing to sing as part of the default state of the world. This is because breathing is more obviously constitutive of our natural state, and thus is more obviously part of the default state of the world, than singing. Again, my moral intuitions follow a similar pattern: continuing to breathe seems to be *more* on the okay side than continuing to sing, even if neither is clearly permissible.

---

<sup>14</sup> The fact that the ordinary killing/letting die distinction seems to be compatible with naturalism about human agency is important. In particular, it blocks a serious objection to the moral significance of the distinction: the charge of anti-naturalism that, e.g., Howard-Snyder raises to Donagan's account of the distinction in terms of the "course of nature" (see Donagan (1977) and Howard-Snyder (2002)). In addition to naturalism, a related reason why the distinction between killing and letting die fails to yield a clear verdict in some cases is that it is unclear to what extent it lines up with the action/omission distinction. There is good reason to believe that the two distinctions don't completely coincide (many people have argued for this), but the boundaries and areas of overlap between the two distinctions are not clear.

Importantly, it seems to me that, not only does the moral indecision about these cases match up with the indecision about the concepts of killing and letting die; our moral intuitions seem to be unclear, moreover, precisely *because* of the indecision about killing and letting die. The reason, for example, that I don't know whether I should keep singing is that I don't know under which circumstances I would be killing someone. In other words, although I want to do the right thing and it seems to me that this requires not killing anyone, I don't know *how* to avoid killing someone: whether by continuing to sing or by discontinuing my singing. Imagine that I somehow manage to make a decision, say, I decide to keep singing. I would then worry about whether I made the right choice, in particular, I would worry that I might have killed Y by doing so. And, again, I am more likely to worry that I might have killed Y in the singing variant of the case and less likely to worry in the breathing variant, since it's more tempting to think that Y was already under the scope of the threat in the breathing variant than in the singing variant. All of this suggests that the moral indecision is *grounded in* the indecision about intervening and failing to intervene. Thus it is further evidence of the existence of a phenomenon of moral inertia.

### **2.3 Deflection's role**

So far we have looked at one manifestation of the phenomenon of moral inertia. This first manifestation involved deflection scenarios of a particular kind: one where a preexisting process could be deflected in such a way that the same number of deaths would result from the deflection, although the identity of the people who die would be different. But, what exactly is the role that deflection plays in giving rise to this form of moral inertia? Does it

play an important role, or would a different kind of scenario have done just as well? In what follows I argue that deflection plays a key, twofold role in giving rise to moral inertia.

First, deflection plays a role in that it provides ideal “test cases” for the moral significance of the intervening/failing to intervene distinction, and thus for the existence of moral inertia. For the relevant deflection scenarios seem to meet the following two conditions simultaneously (whereas most arguments for the moral significance of the distinction fail on at least one of these counts): (a) our moral intuitions in these cases are very strong, and (b) it is easy to imagine that the cases are “equalized” with respect to other potentially morally relevant factors. The fact that the scenarios we looked at are scenarios of deflection facilitates their meeting both of these conditions simultaneously. Let me explain.

Regarding (a), the question you need to ask yourself in a deflection scenario is which of two acts, deflecting the process or failing to deflect it, you ought to do, or it is permissible for you to do, when those are the only alternatives. And it seems generally easier to answer this kind of question than a question of the type: “Which of two acts, a1 and a2, performed by two people, is morally worse (or morally better)?”, which might require comparing acts with a similar moral status (e.g., acts that are both very bad). In the literature on this topic, the question of whether there is a moral difference between killing and letting die tends to be put in the form of a comparative question, e.g., the question of whether it’s worse to make someone drown in a bathtub than to “merely” let him drown (Rachels (1975)). It is particularly hard to answer a question of this type. For both acts seem outrageous. So maybe

they are both equally bad. Or, then, again, maybe they are not: maybe one is worse than the other, but it's hard to see the difference precisely *because* they are both so bad.<sup>15</sup>

Now, deflection scenarios are clearly not the only scenarios where a choice is made between two alternative courses of action. So this isn't enough to single them out as unique. But consider (b). Deflection cases do seem to be special in securing (b). For it is particularly easy to imagine that deflection cases are equalized with respect to other morally relevant factors. In particular, it's much easier to imagine that deflection cases are thus equalized than it is to imagine that cases involving *the launching of (new) threats* are. The examples of killing that are commonly offered in the literature are examples involving the launching of threats; these are contrasted with failures to stop (old) threats, which are cases of letting die. It is hard to imagine cases of this type that are sufficiently equalized with respect to other morally significant factors.

Consider, for instance, the agent's intention: the launching of a threat typically involves a bad intention, and an intention that is much worse than that involved in the mere failure to stop an existing threat. By contrast, the deflection of an existing threat doesn't typically involve a bad intention, or an intention that is much worse than that involved in the failure to deflect a threat. On the contrary, we typically deflect threats in order to save people's lives.<sup>16</sup> Also, the launching of a threat typically carries with it a higher certainty that harm will ensue than the mere failure to stop an ongoing threat. In contrast, the deflection of

---

<sup>15</sup> See, e.g., Kagan (1998), p.99. Certainly, there are other cases of killing and letting die for which the moral difference is much more obvious (say, sending poisoned food to starving children in developing countries versus not sending them aid of any sort), but in those cases it's less clear that other morally significant factors are sufficiently "equalized" (and thus that condition (b) above is met). The same seems to be true of Foot's famous "Rescue I" and "Rescue II" examples (see Foot (1984)). In her (1989) Malm calls cases like Runaway Train "simple conflict examples," which she distinguishes from "comparison examples."

<sup>16</sup> Moreover, why should the deflection of the threat from X to Y suggest a bad intention on the agent's part? Why assume that the agent deflected the threat because he wanted to kill Y, not because he wanted to save X?

an ongoing threat doesn't typically carry with it a higher certainty that harm will ensue than the failure to deflect a threat. Again, on the contrary, the deflection of threats typically makes the occurrence of harm less likely. Finally, it's typically much more costly to stop an ongoing threat than to fail to start a (new) threat. By contrast, it isn't typically much more costly to deflect an already existing threat than to fail to deflect it. Granted, in this case we might still see a small difference: it might be slightly more costly to deflect a preexisting threat than to fail to deflect it, at least generally. But the point remains that the difference in cost between deflecting a threat and failing to deflect it is generally smaller than that between stopping a threat and failing to start one. For altogether stopping a threat typically requires more effort than merely deflecting one.

As a result, deflection cases are good test cases for the existence of moral inertia, in particular, they are much better cases than the ones that are usually offered in the literature. This is due to the existence, in deflection cases, of a process already in motion (such as a threat) whose course could be altered in certain ways. In these cases, neither the agent's intervening nor the agent's failing to intervene are regarded as the *launching* of a threat, and thus it is much easier to imagine that the scenario where the agent intervenes and the scenario where the agent fails to intervene are on a par with respect to other morally significant factors. The first role that deflection plays in the case for moral inertia, then, is that it allows us to isolate the moral significance of the intervening/failing to intervene distinction from that of other potentially morally significant factors.

The second role that deflection plays is that, without the preexisting process already "in motion" present in deflection cases, we would be at a loss identifying the "default" state of the world, and thus there would be no moral inertia. As we will see, in cases where this

deflection structure is not present, or where it is less obviously present, we see a corresponding change in our moral intuitions: either we tend to have different intuitions, or our intuitions become much less clear.

Consider, for example, a case devised by Michael Tooley: the famous “diabolical machine” case.<sup>17</sup> We are told that two children, Mary and John, are trapped inside a machine. One child will die regardless of what we do, but we can decide who lives and who dies by pushing a button or failing to push it: if we push the button, John will be killed; otherwise Mary will be killed. This is all we are told. Tooley claims that it’s not the case that we ought to refrain from pushing the button (he claims that, if possible, we should flip a coin; otherwise it doesn’t matter what we do).

What is our reaction to this case? I confess that my initial reaction is not as strong as that concerning Runaway Train and Floating Drug (although I think I wouldn’t push the button). But I suspect that the reason my reaction is less strong in Tooley’s case is that the case is under-described in important ways. In particular, Tooley doesn’t tell us what the inner mechanisms of the machine are (only that pushing a button would result in a death and failing to push it would result in another). As a result, we are less likely to see the case as a case of *deflection*. I submit that, were we to see it as such, our reaction would be just as strong as in Runaway Train.

First, notice that, for all Tooley’s said, it might be that this is how the machine works. There is a detector that would be triggered by the failure to push the button at a given time,  $t$ , and a different detector that would be triggered by the pushing of the button at  $t$ . The triggering of each detector starts a process leading to one child’s death. Say, the first

---

<sup>17</sup> Tooley (1994), p. 108.

mechanism shoots out a pink bullet that kills Mary, and the second mechanism shoots out a blue bullet that kills John. If this is how the machine works, one might be tempted to regard my behavior in either case as *starting* a deadly process: I start one process (the pink-bullet process) if I fail to push the button at *t* and I start another process (the blue-bullet process) if I push the button at *t*. In other words, in these circumstances, we don't clearly perceive Mary as *already* threatened by a process that is already "in motion". This is why, I suggest, we don't have as clear an intuition that we ought not push the button.<sup>18</sup>

But now imagine that this is how the machine works: there is a salient process already in motion that is threatening to kill Mary, say, a bullet has already been launched in her direction, and pushing the button would deflect *it*—that same bullet—towards John. My intuition is much stronger in this case, and it lines up with that about Runaway Train: it seems to me that I ought not push the button. Again, I would think that John got lucky and Mary unlucky, and that I don't have enough reason to intervene and change this fact. The reason we see Mary as being unlucky in this case is the existence of a salient process already threatening her. In other words, when the deflection structure is clear, our intuitions clearly support the principle of moral inertia.<sup>19, 20</sup>

---

<sup>18</sup> Note that it's still true that Mary would have died if I hadn't pushed the button. This is a reason to think that deflection scenarios cannot be characterized in purely counterfactual terms, as Malm attempts to do in her (1989), n.11.

<sup>19</sup> Intervening in the following way is also impermissible: a bullet has already been launched towards Mary; one could launch a *new* bullet towards John, which would kill John and also deflect Mary's bullet away from its path. Although in this case Mary is already under the scope of a threat, we kill John by launching a new threat onto him, not by deflecting the old threat. So this isn't a deflection case. But I think that cases of this kind are less convincing as a refutation of Tooley and in support of moral inertia. For, as I pointed out earlier, starting a threat typically carries with it a worse intention than merely failing to stop a threat. Hence Tooley could say that the reason it seems morally wrong to intervene in this case is that we can't help but see a bad intention underlying the act.

<sup>20</sup> Another interesting example to consider is a third-person "ducking" scenario (Boorse and Sorensen (1988), p. 118). Boorse and Sorensen seem to believe that, just as it would be okay for X to duck a bullet even if he realizes that Y (who was standing in queue behind him) will get hurt, it would be okay for a bystander to push X out of the way in the same circumstances (even if he realizes that Y will get hurt instead of X). This strikes me as

There is yet another way in which Tooley's diabolical machine case is under-described, and which might also account for our diverging moral intuitions in that case. I discuss this in the next section. At the same time, as we will see, reflecting on this point can also help us get a better grasp of the phenomenon of moral inertia on this first manifestation.

## 2.4 Wrongdoing

One cannot help but wonder: in Tooley's imaginary scenario, why is it that the children are trapped in the diabolical machine in the first place? It is natural to assume foul play, e.g., it is natural to assume that some diabolical agent put them in there. But, at least in other contexts of a similar type, wrongdoing can affect what we think we can permissibly do. A clear example of this is: imagine that Y tied X to the left-hand track before getting himself tangled up in the right-hand track. It is clear that in this case I may deflect the train to Y. For the train is only threatening to kill X because of Y's wrongful conduct. Or imagine that someone flips the switch in Runaway Train (thus acting wrongly). Then maybe it's okay for me to "undo the injustice" and flip the switch again to return it to its original state. In other words, what we regard as the "inertial" state doesn't seem to be determined by other agent's acts *to the extent that they involve wrongdoing* (or, at least, to the extent that they involve wrongdoing of some kind; more on this later).

---

wrong, if other things are really equal, e.g., if we know that Y won't be able to escape the threat in the meantime. My impression is that, although it would be okay for X to duck the threat himself, it wouldn't be okay for a *bystander* to push him out of the way if other things are equal, just like, although it would be okay for X to redirect the train away from himself and towards Y (if he had a remote control that he could use to redirect the train), it wouldn't be okay for a bystander to do the same thing. An additional complication is that, in a ducking case, it seems less clear that we regard the bystander's act as a *deflection* of a threat. For he is not directly acting on the *threat* itself (as in Runaway Train), but on X. This might explain our perceiving the ducking cases in a slightly different way.

This consideration can also help us address an objection to the view that there is moral inertia, raised by Judith Thomson. In the context of her discussion of the “trolley problem,” Thomson wrote:

There is no Principle of Moral Inertia: there is no prima facie duty to refrain from interfering with existing states of affairs just because they are existing states of affairs. A burglar whose burgling we interfere with cannot say that since, but for our interference, he would have got the goods, he had a claim on them; it is not as if we weigh the burglar’s claim on the goods against the owner’s claim on them, and find the owner’s claim weightier, and therefore interfere—the burglar has no claim on the goods to be weighed. (Thomson (1976), p. 84)

In normal circumstances, Thomson is right about the burglar: the burglar has *no* claim on the goods, not even one that is easily outweighed by other people’s claims. Still, this case is very different from, say, Runaway Train. In Runaway Train, the person on the left-hand track just happens to find himself in the path of the train, as a matter of mere “chance,” through no one’s fault. The burglar, by contrast, does not just “happen” to find himself near the goods: crucially, he’s near the goods *because* of his wrongful conduct. As a result, we should not expect moral inertia to kick in in this case, just like we wouldn’t expect it to kick in in other cases of wrongdoing.

Now, does this mean that wrongful behavior can never contribute towards what we regard as the “inertial” state of the world? Or perhaps some kinds of wrongdoings can? (If so, what kinds?) This is another important source of unclarity in our intuitions. Imagine that a

child who was playing by the tracks early this morning, when he should have been in school, flipped the switch and, as a result, the train is now threatening Y instead of X. Were it not for the child's wrongful behavior (his missing school without permission), Y wouldn't now be under the scope of the threat. Is it permissible to flip the switch in this case (that is, to return it to its original position)? I'm not sure. On the one hand, the child shouldn't have been there in the first place, so there is a sense in which Y should never have been threatened by the train. But, on the other hand, the fault in question is small, maybe small enough that we would be tempted to disregard it. Moreover, the flipping of the switch itself was completely innocent; what is faulty, we are assuming, was the child's being in that location at that time (a necessary condition for the flipping of the switch). This suggests that the type of fault and the seriousness of the fault can make a difference to whether we regard the wrongdoing as contributing to the inertial state of the world. Another factor that presumably matters is time: if the fault took place a long time ago, we'll probably be more tempted to disregard it than if it took place a few seconds ago.<sup>21</sup>

At any rate, I think it is clear that we regard wrongdoing as a special case. So, with this in mind, we are now in a position to offer a more precise formulation of the principle of moral inertia, on its first manifestation. Let P be any threatening or benefiting process that meets the following condition: it is either a "natural" process (a process that is not the result of human behavior in any way), or a process that is also the result of human behavior, but of a morally innocent kind), or, at most, a process that is also the result of some limited class of

---

<sup>21</sup> So what should we think about the version of Tooley's diabolical machine case where an evil man is responsible for the setup? I'm not really sure. I certainly see why it's more tempting to say in this case that it would be permissible to flip a coin. But, still, I'm not sure that this is justified.

wrongful human behaviors. We can then spell out the first manifestation of moral inertia as the following principle of non-intervention:

*Moral Inertia (Non-intervention):* Given any two moral agents (or two equinumerous groups of moral agents) X and Y, if X is originally under P's scope and Y isn't, and if you could intervene so that P is deflected from X to Y, then, other things being equal, you ought not intervene.<sup>22</sup>

This principle summarizes the first source of moral pressure for failing to intervene, or for leaving things "as is." In the next section I examine the second source.

### **3. Second manifestation: Constraints on interventions**

Consider the following variant of Runaway Train:

*The three-track variant:* This time there are five people stuck on the main track. The switch is originally set for that track. There are two side tracks, A and B, containing three people and one person, respectively. If you do nothing, 5 will die; if you deflect the train to A, 3 will die; if you deflect it to B, only 1 will die.<sup>23</sup>

---

<sup>22</sup> Perhaps we should add to the antecedent of the principle: "and P is the only major threatening or benefiting process of its kind in the situation." For imagine that a train T1 is threatening X and another similar train T2 is threatening Y, and that you could deflect T1 to Y and T2 to X. Intervening doesn't seem wrong in this case, if all other things are equal. Thanks to Ryan Wasserman for raising this point.

<sup>23</sup> I am indebted to Stephen Yablo for bringing up this kind of scenario in conversation a few years ago.

I take it that most people's intuitions are as follows: other things being equal, you don't have an obligation to intervene, but, if you do intervene, then you ought to intervene in a specific way: you ought to deflect the train to track B, and not to track A. In other words, although it wouldn't be wrong to fail to deflect the train, or to deflect it to prevent the larger number of deaths possible, it would be wrong to deflect it in a way that doesn't result in the largest number of lives being saved.

Again, a way to see the strength of this intuition is by examining our potential reasons for intervening and not intervening, as well as other people's potential grounds for complaint. What could be your reason for not intervening? Presumably, that you don't want to kill anyone (by failing to intervene, you don't kill anyone: you only let the five die). What could be your reason for switching to B? Presumably, that you want to minimize the number of deaths. Now, what could be your reason for switching to A? It seems that there couldn't be any good reason to do so. In fact, it seems that those three people who would die (and their relatives) would have legitimate grounds for complaint: if you divert the train to A, you kill those three people, while not minimizing the overall amount of harm.

We may imagine a similar variant of the Floating Drug case:

*The three-group variant:* The drug is headed towards one person, who would need the whole dose to survive. There are two other groups of people in the water: a group of five (each would need one-fifth of the drug) and a group of three (each would need one-third of the drug). You could redirect the drug to either group.

Similar remarks apply to this case: presumably, other things being equal, you don't have an obligation to intervene but, if you do, you must act so that the largest number of lives are saved (you must avoid deflecting the drug to the group of three).

Again, this suggests that there is moral pressure to leave things "as is": it is easier to act impermissibly by intervening than by failing to intervene. In this case, this is because, if you intervene, you have to be careful about *how* you intervene. We may express this idea in the form of a new principle. Let P be any process that satisfies the conditions described before (in section 2.4). Then the principle says:

*Moral Inertia (Constrained intervention)*: If a group of moral agents X is under P's scope and you could intervene in more than one way, by deflecting P to one out of two or more groups of people, there are more moral constraints on intervening than on failing to intervene.

For example, if P is a threat and the groups have different number of people in them, as in the three-track variant of Runaway Train, then (if all other things are equal) the only permissible way of intervening is to deflect P so as to minimize the number of deaths. There is no similar restriction on failing to intervene (you may fail to intervene even if this doesn't minimize the number of deaths). Or imagine that all the groups are equinumerous. Then, if the people in one of the groups that you could deflect P to would die a quick and painless death but the people in the other groups wouldn't (and all other things are equal), then the only permissible way of intervening is to deflect P to the people who would suffer less. Again, presumably, there is no similar restriction on failing to intervene (it is okay to fail to intervene even if the

people on the main track would suffer a worse death). And so on. In general, the principle says that interventions are subject to more constraints than failures to intervene: you would need more reason to intervene than to fail to intervene.

Once again, it is natural to wonder about the connection between moral inertia (on this second manifestation) and the debate concerning the moral significance of the killing/letting die distinction. In this case it is less clear what the connection is. The thesis that the killing/letting die distinction is morally significant is sometimes understood as the thesis that, other things being equal, killing is worse than letting die.<sup>24</sup> If the thesis is understood in this way, then it doesn't apply to cases like the three-track variant of Runaway Train, for other things are not equal in the different scenarios where you kill and let die (different numbers of people die in each scenario). Still, it seems that we could understand the thesis of the moral significance of the distinction more broadly. For example, we could take it to be the claim that letting die is "more easily permissible" than killing, in the sense that, if you kill, certain conditions have to obtain for your act to be permissible, which don't necessarily have to obtain for your act of letting die to be permissible. I don't have any objection to understanding the thesis about killing and letting die in this way. It is clear that, understood in this way, the thesis about killing and letting die is, again, intimately connected to the thesis of moral inertia.

Note that, here too, our moral judgment hinges on our judgment about when the agent would be intervening and when he would be failing to intervene. In particular, in scenarios where it's unclear under what conditions you would be intervening and under what conditions you would be failing to intervene, it's also unclear (and to a similar degree) what moral constraints there are on your acting in certain ways. Consider, for instance:

---

<sup>24</sup> See, e.g., Rachels (1975) and Thomson (1976).

*The singing variant of the three-track variant:* You are singing a tune when you approach the switch. You see that the mechanism that switches the train to A is equipped with a sound sensor. If you keep singing, the three people on track A will die; if you stop singing, the five people on the main track will die. In both cases, this is assuming that you don't pull a lever that switches the train to B; if you pull the lever, the one person on track B will die instead.

In this case it's unclear that it's wrong to act in such a way that the three people on track A die (that is, to simply continue to sing). This is, I suggest, for the same reasons as before: because it's unclear what we regard as the default state. Given that you were already singing when you approached the switch, there is more of a temptation in this case to regard the train's turning to A as the default state; hence, there is less of a temptation to think that acting in a way that results in the train's turning to A is wrong. But there are also reasons to regard the train's continuing on the main track as the default state; after all, the switch is originally set for the main track. Hence, it's unclear when you would be intervening and when you would be failing to intervene, and this explains the unclarity of our moral intuitions. (Here, too, I presume that there would be an even *stronger* temptation to regard the train's turning to A as the default state if the A-switch were triggered by your breathing instead of by your singing. In a breathing variant of the case, it seems less clearly wrong to fail to hold your breath, even if whether it's ultimately permissible to continue breathing is not totally clear.)

In the previous section we presented a case for moral inertia on the basis of deflection scenarios of a certain kind. As we saw then, the role played by deflection in those scenarios

was key. Here, too, the role played by deflection is crucial. For, interestingly, there is no similar argument for restrictions on interventions based on non-deflection scenarios. Imagine that you are thinking of donating part of your liver. This is clearly supererogatory: you have no obligation to do this (even if you don't have to sacrifice your life to do this: the liver can regenerate to full size in a matter of weeks). However, imagine that there are two people who could both benefit from the transplant (say, each would need only a small portion of it). Alternatively, you could give it to someone (just one person) who needs a larger portion to survive. It seems that in this case there are no moral limitations on how you can intervene; in particular, it would be permissible for you to help only one person, if this is what you desire. After all, it's your liver, and you can do whatever you want to do with it (as long as it's for a good cause). Or at least this is what ordinary morality would say: it's permissible for you not to help at all, and it's also permissible for you to help as much as you want, or however you choose. I suggest that this is because in this case you wouldn't be diverting a preexistent train of events (either potentially harming or benefiting) from a group of people onto a different group of people. If you decide to help, you are starting a *new* process that would potentially benefit people. Given that no one is originally under the scope of a potentially harming or benefiting process, moral inertia doesn't kick in.

#### **4. Conclusion: Deflection revisited**

I conclude that, according to ordinary morality, there is moral inertia. Moral inertia arises for a specific class of interventions: those where an agent interferes with a process that can clearly be seen as already "in motion," and one where there is a path that can clearly be seen as the "preset" path. Under those conditions, there is moral pressure against deflecting those

processes from one group of people to another. This moral pressure takes the form of strict prohibitions on such deflections, or of constraints on such deflections.

But is moral inertia defensible? Is there a rational justification of the moral pressure to fail to intervene? Whereas I cannot give this question full consideration here, I will outline one main objection that could be raised against the defensibility of moral inertia—in particular, one that attacks the intelligibility of the concept of deflection (which was my focus here)—and I will explain how the objection could be addressed.<sup>25</sup>

As I pointed out, we feel the pressure of moral inertia when we identify a process that is already “in motion.” But, the objection goes, whether we see a process as such is something that seems to depend on our specific psychological features (our beliefs, desires, expectations, etc.). What’s salient to you might not be salient to me, and even if the same processes were equally salient to everyone in some cases, it could be because of some general feature of human psychology, not because of a true feature of the world. So the worry is that the concept of deflection on which moral inertia rests might be too subjective to be able to carry any genuine moral weight. If so, moral inertia would be indefensible.

Can this objection be successfully answered? My impression is that this is an issue that needs to be settled in conjunction with other questions about deflection. In the first section of this paper I pointed out that deflection scenarios are thought (by some people, at least) to have implications for the metaphysics of causation. In particular, it is natural to see a causal difference between certain deflection scenarios (“switches,” as they are sometimes

---

<sup>25</sup> An important objection that I’ll bypass, and that I think needs to be addressed, is this: Assuming that we can make sense of the distinction between intervening and failing to intervene, why think that this distinction might have moral significance? It is tempting to think that it’s because, when an agent fails to intervene with some ongoing state of affairs, the outcome is not “due to the agent” in any way. But this is misguided, because in all the scenarios we have looked at, the agent’s failure to intervene still helps to determine the outcome. (Bennett raises this type of objection against Donagan’s views on killing and letting die in Bennett (1995), ch.7.)

called) and other scenarios (notably, “preemption” scenarios). On the assumption that there is such a causal difference, the prospects of giving an objective account of the concept of deflection, or of the concept of being under the scope of a threat, seem good.<sup>26</sup>

The guiding idea behind such an objective account of deflection would be this. Deflection scenarios are cases of interference with a preexisting threat; these are contrasted mainly with cases where there is no preexisting threat, or where there is a preexisting threat but a new threat is launched. As explained in section 1, in cases where the deflection doesn’t affect the outcome, this contrast seems to be a causal contrast: in those cases, deflecting an old threat isn’t causing the outcome, but launching a new threat is. For example, temporarily diverting a missile, which then resumes its original path, is not causing the destruction of the target. By contrast, launching a new missile that hits the target before the old one does is causing the destruction of the target.

Call those deflection cases the “simple” cases. The simple cases are accounted for in causal terms, or in terms of whatever makes the causal facts true. On this basis, we could then hope to characterize other deflection cases (such as Runaway Train) as those cases that have certain features in common with the simple cases. The challenge would be, of course, to identify the features that do the trick. But, assuming that the simple cases can be characterized causally, this looks like a promising place to start.<sup>27</sup>

---

<sup>26</sup> A causal difference is not the only kind of structural difference that could fill the bill, but it is one obvious candidate. Another option would be to try to distinguish deflection scenarios from other types of scenario in terms of the different kinds of processes that obtain in them.

<sup>27</sup> This is assuming that we regard causation as an objective relation (one that obtains between events in the world independently of our subjective or inter-subjective experience). Ned Hall recently argued for a view of causation according to which, interestingly, causation rests on the distinction between *default* states of the world and *deviant* states (departures from the default states). For Hall, this is because causation rests on counterfactual dependence and counterfactual dependence, in turn, rests on the default/deviant distinction. According to Hall, although typically there is a natural and intuitive way to draw the default/deviant distinction, the distinction is not purely objective. As a result, causation doesn’t turn out to be fully objective either. (See Hall (2007),

## References

- Bennett, Jonathan (1995) *The Act Itself*, New York: Oxford U.P.
- Boorse, Christopher and Roy Sorensen (1988) "Ducking Harm," *Journal of Philosophy* 85, 3: 115-134.
- Donagan, Alan (1977) *The Theory of Morality*, Chicago: The University of Chicago Press.
- Hall, Ned (2007) "Structural Equations and Causation," *Philosophical Studies* 132: 109-36.
- Howard-Snyder, Frances (2002) "Doing vs. Allowing Harm," *Stanford Encyclopedia of Philosophy*.
- Foot, Philippa (1984), reprinted in Steinbock and Norcross, *Killing and Letting Die*, New York: Fordham University Press, second edition, pp. 280-289.
- Kagan, Shelly (1998) *Normative Ethics*, Boulder: Westview Press.
- Malm, Heidi M. (1989) "Killing, Letting Die, and Simple Conflicts," *Philosophy and Public Affairs* 18, 3: 238-58.
- Maudlin, Tim (2004) "Causation, Counterfactuals, and the Third Factor," in Collins, Hall, and Paul (eds.), *Causation and Counterfactuals*, Cambridge, Mass: M.I.T. Press, pp. 419-443.
- Paul, L. (2000) "Aspect Causation," *Journal of Philosophy* 97, 4: 235-56.
- Rachels, James (1975) "Active and Passive Euthanasia," *New England Journal of Medicine* 292, 9: 78-80.
- Sartorio, Carolina (2005) "Causes as Difference-Makers," *Philosophical Studies* 123, 1-2: 71-96.

---

especially pp. 126-7. See also Maudlin (2004), which inspired Hall's view.) If this view had any merit, it would be an obstacle for the project I have described.

Thomson, Judith (1976) "Killing, Letting Die, and the Trolley Problem," *The Monist*,  
reprinted in *Rights, Restitution, and Risk*, Cambridge: Harvard University Press, 1986: 78-93.

Tooley, Michael (1994) "An Irrelevant Consideration: Killing versus Letting Die," in  
Steinbock and Norcross, *Killing and Letting Die*, New York: Fordham University Press,  
second edition, pp. 103-111.

Yablo, Stephen (2002) "De Facto Dependence," *Journal of Philosophy* 99, 3: 130-48.