Philosophia (2008) 36:1–15
DOI 10.1007/s11406-007-9099-z

Damned If You Do; Damned If You Don’t!

Frances Howard-Snyder

Received: 28 June 2007 / Revised: 6 September 2007 / Accepted: 10 September 2007 /
Published online: 13 November 2007

Abstract This paper discusses the Principle of Normative Invariance: ‘An action’s moral status does not depend on whether or not it is performed.’ I show the importance of this principle for arguments regarding actualism and other variations on the person-affecting restriction, discuss and rebut arguments in favor of the principle, and then discuss five counterexamples to it. I conclude that the principle as it stands is false; and that if it is modified to avoid the counterexamples, it is gutted of any interest or power.

Keywords Moral responsibility · Future people · Ought · Principle of Normative Invariance · Actualism

My Youthful Mistakes Redeemed

Watching new students milling around Red Square at my university the day before classes start, as they sign up for political organizations, social groups, academic opportunities, I remember my first days at university. I recall my youthful idealism, and regret the way I squandered my freshman and sophomore years. I should have studied harder; I shouldn’t have drunk so much.... I could go on and on.

Given widely accepted metaphysical assumptions, however, if I had done things differently, my sons wouldn’t exist.[1] None of my mistakes are so terrible that their bad results outweigh the good of these two children. So I don’t regret my “mistakes” — all things considered. Given my values, worlds without these mistakes (or at least the closest worlds without these mistakes) are worse than the actual world. So, it was good and right that I made these choices.

Of course, had I not made these mistakes, I would have had other children. (Let’s suppose I would have procreated.) I would value them as much as I value my actual

[1] As Josh Parsons points out, one doesn’t have to accept Kripkean necessity of genetic origins in order to accept this just that if I had had a different spouse or even waited a month or two to conceive, whatever other children I would have had would have been the equivalent of half or full siblings of my actual children. See Parsons, “Why the Handicapped Child Case is Hard,” *Philosophical Studies*, 112. 2, pp.147–162, 2003.

F. Howard-Snyder (✉)
Department of Philosophy, Western Washington University, 516 High Street, Bellingham, WA 98225, USA
e-mail: Frances.Howard-Snyder@wwu.edu

children, and I would be glad that I made the choices I made. In that world, I would say, “It was right and good that I didn’t make the mistakes.”

Of course, the fact that my actual self and my counterfactual self disagree is not very surprising. What *would be* surprising, however, is if my actual self and my counterfactual self disagreed and *were both right.* What if we were? What if my mistakes were right; but also, that if I hadn’t made them, they would have been wrong?

Some will say that this is silly. Either my actual children are so fabulous that they are indeed better than the children I might have had, and so, their perfection does justify my mistakes no matter which world we consider the matter from; or else, more likely, they are really no better than some or all of the counterfactual children, in which case, their relative goodness does not justify my mistakes. Either way, the puzzle evaporates.

Well, not exactly. This response focuses solely on what we might call “purely objective” value; and it’s not obvious that that is the only sort of value there is. Most of us would regard a world where our children or spouses were annihilated and replaced by perfect duplicates as an inferior world, even though the overall amount of happiness is the same.[2] And this seems a reasonable preference, one we can endorse from a third person perspective. Moreover, if someone were faced with a choice of bringing about this duplication, it would seem wrong to do it even if she could produce some small extra good in the process - e.g., the duplicate would live slightly longer.

Now, consider a world where the duplication takes place from the very beginning, prior to the beloved’s conception. It is better than a world where the beloved is *annihilated* and replaced, but it still seems an inferior world. Hard-nosed utilitarians will express impatience at this point. But it is a widespread thought shared by philosophers and non-philosophers alike.[3]

Analogy with Actualism

Let’s broaden this discussion to connect with philosophically significant issues. A number of philosophers have recently discussed or defended the following principle or some slight variant on it:

> The actuality principle: One state of affairs is at least as good as another if and only if it is at least as good for those people who *actually* exist.[4]

[2] To be clear, I’m asking you to think about a world in which your family is annihilated and replaced, and you are completely unaware in that world of these events. It seems that looking at such a world from the outside we can recognize it as inferior, even if we wouldn’t complain if we were in the world.

[3] Not all non-utilitarians share it, of course, but the fact that it is widespread makes it at least worth considering. Does one have to embrace some sort of agent-centered approach to morality to see this? I’m not sure, but even if one does, it is still a view that cannot be rejected out of hand.

[4] See Josh Parsons, “Axiological Actualism,” *Australian Journal of Philosophy* vol. 80, No.2, June 2002, and Gustav Arrhenius, in a number of places, including “The Person Affecting Restriction, Comparativism, and the Moral Status of Potential People”, Ethical Perspectives, no. 3–4, 2003, discusses principles close to this. This is related to an idea defended by Jan Narveson in, “Utilitarianism and New Generations,” *Mind* 76 (1967). Incidentally, the spirit of this principle would be preserved if we replaced the word “people” with the word “individual” to accommodate non-human animals. It has been brought to my attention that this principle might rule out the value of unseen beautiful paintings or landscapes. Maybe so and maybe that is a reason to reject the principle, or maybe it can be adjusted to avoid that implication. I am not interested in defending the principle against all objections here, however.

Imagine a big, happy family, the Huxtables, trying to decide whether to have another child. If they do, they reason correctly, the child will be happy, but each member of the rest of the family will be slightly worse off. We might even suppose (if this is coherent) that the total amount of happiness or other good the new child will enjoy will be greater than the sum of what will be lost by the other family members. They decide not to have the child. According to the actuality principle, the result was better than its alternative. The comparison between the happiness of the actual family members in the two scenarios is what matters morally. The hypothetical happiness of the non-actual child is irrelevant.

But now change the story. Suppose that the Huxtables had decided to have a child and that they did so. Now, this child, call her Rudy, is actual. Her happiness is sufficiently great to outweigh the sum of all the small losses in well-being each of the family members sustains. Now her well-being does count.[5] If that happened, then it would be better to have the child.[6]

Now consider a normative variant on the actuality principle: An act is *right* iff its consequences are at least as good for those people who actually exist as are the consequences of the act's alternatives. In that case, if the Huxtables do not have a child, it is right for them not to have a child (and having a child would be wrong); whereas, if they have the child, then it is wrong for them not to do so. That looks pretty paradoxical. (Another normative variant on actualism simply says that an act is at least as good on the oughtness scale i.e., is at least as good *an act* iff it is at least as good for those people who actually exist as all alternatives.[7]) On this variant, we can accommodate the intuitive reaction that it is perfectly permissible to conceive Rudy and permissible not to conceive Rudy, but add that conceiving Rudy is a better action than not doing so in a world in which she is conceived, and failing to conceive her is a better action in a world in which she is not conceived. This is still paradoxical.

We could also imagine this on a grander scale. A national government with a declining population is considering a proposal to expand the size of its country's population, perhaps by offering tax incentives to bear children. If the results parallel the Huxtable family results in terms of benefits to existing people versus benefits to new people, then we get similar results. Suppose the extra population will be fortunate enough to offset the losses to the already existing population. In that case, if the government does choose the policy that leads to the births of these extra people, that policy is the right one; whereas, if it rejects that policy, then the policy

[5] Note that, on this view, it counts at all times, including times prior to Rudy's existence. What matters is that she is actual, not that she currently exists. There is a small complication here. Some people say that we cannot compare Rudy's well-being in the actual world with her well-being in the world where she does not exist. Josh Parsons gives an elegant argument for thinking that we can in "Axiological Actualism" p.144. I shall assume that such comparisons make sense. If we cannot make such comparisons, the Actuality Principle generates other paradoxical results of the sort I am interested in. So, nothing really hangs on these comparisons.

[6] John Broome discusses an example like this. See *Weighing Lives* (Oxford: Oxford University Press, 2004) p9 as does Gregory Kavka in "The Futurity Problem" in Obligations to Future Generations, Sikora and Barry (ed.) (Philadelphia, Temple University Press: 1978).

[7] To think about what this means, consider two wrong acts, one of which is less wrong than the other; or two permissible acts, one of which, that we might call supererogatory, we have more moral reason to perform than the other.

would have been the wrong one.[8] Although there is something attractive about this idea that actual people matter (more), it faces a formidable objection.

Objection: The Principle of Normative Invariance

Erik Carlson formulates the principle of Normative Invariance,

> “NI: An action’s moral status does not depend on whether or not it is performed.”[9]

This principle would imply the denial of the Actuality Principle. Since, if it is true, then which actions are right will depend on which people actually exist, and in some cases, the relevant action will be the action of making it the case that some people exist.

Should we accept it? What can be said for and against it? Well, it does rule out a couple of things we have independent reasons to rule out. First, it rules out a kind of crude Divine Command Theory or individual ethical relativism according to which an agent automatically does the right thing because whatever he or she chooses to do turns out to be the right thing in virtue of having been chosen. But there are other, better, reasons to object to such views. Second, it rules out a case where an agent cannot help but do wrong, which seems to be in tension with (at least the spirit of) “ought” implies “can” (OIC). Consider a case where an agent is faced with two options, A and not-A, where, given that she does A, A is wrong; but where, if she had done not-A, A would have been right and not-A wrong. That result seems to violate something close to OIC. Although it doesn’t strictly speaking violate OIC, since she *can* do A and she *can* do not-A, and one of these is such that she ought to do it, that is to say, that doing what she ought to do (*de re*) is something within her power, it is nevertheless not within her power to do what she ought *de dicto*. It is not possible for her to satisfy the description, “doing what she ought” or to avoid the accurate description “having done wrong.” The kind of arguments that support OIC (especially those having to do with fairness and practicality) seem to rule out a moral theory that leaves an agent in this sort of bind.[10] So, NI entails something true. But that doesn’t mean that NI is itself true.

Perhaps it would be *ad hoc* to rule out cases where one acts *wrongly* no matter what one does, while allowing cases where one *does what one ought* no matter what one does. But if this is inconsistent that must be because the inverse of OIC “ought

[8] At this point, it may look as if the Repugnant Conclusion made famous by Derek Parfit in *Reasons and Persons* (Oxford: Clarendon Press, 1984) rears its repulsive head. If the authorities decide to keep adding people to the world, we’ll end up with a world with tens of billions of people living just barely tolerable lives, and it will turn out that those decisions were right. Since it’s a well-known fact that many theories, including straightforward utilitarianism, lead to the Repugnant Conclusion, and theories that avoid it lead to different troubling results, let’s leave this concern to one side.

[9] Eric Carlson, *Consequentialism Reconsidered*, p100 (Dordrecht: Kluwer Academic Publishers, 1995). Carlson got the idea and the name for the principle from Wlodek Rabinowicz. Gustav Arrhenius endorses this principle and uses it against Actualism in Arrhenius, Future Generations: a Challenge for Moral Theory, Uppsala University, 2000, as does John Broome, in *Weighing Lives* (Oxford: Oxford University Press, 2004).

[10] See my “’Cannot’ Implies “Not Ought”,” *Philosophical Studies*, 2006.

to x" implies "can avoid x" (OICA) is as plausible as OIC and is motivated by the same sorts of considerations. But OICA is less plausible than OIC. I cannot shoot my best friend (maybe I am psychologically incapable of it or I lack a weapon). Does that mean that it is not true that I ought not to shoot my best friend? Not so obvious. If there were a genuine counterexample to OIC it would be a case of conflicting forces: a moral force pushing the agent to do x, and some other physical or psychological force - preventing her. It is plausible to think that this is a case where the moral force is overridden or cancelled somewhat analogously to the way one moral force can be overridden by another (e.g., the familiar phenomenon of conflicting prima facie duties) or where one physical force is cancelled or overridden by another.

In the case of a counterexample to OICA, we have something that looks more like a case of overdetermination. Morality "prevents" me from shooting my friend. My love for my friend also prevents me. Why not say that each of these is sufficient? And that they do not cancel each other out?[11] Here's Carlson's own argument for the Principle of Normative Invariance:

> I believe that a reasonable theory should... be 'action-guiding' for an agent with complete knowledge of all morally relevant facts in the situation in question. That is, if T is a moral principle, P should be able to use T as a decision-making procedure in S, provided that she knows everything that is relevant, according to T, to what she ought to do in S..... Theories that violate [the Principle of Normative Invariance] do not satisfy this criterion..., since they include facts about what P will do in S among the morally relevant facts. Full knowledge of the relevant facts hence presupposes at least partial knowledge of what P will do in S. P's having such knowledge, however, is incompatible with her making decisions or deliberating about what to do in S. It is conceptually impossible to deliberate about what to do in a certain situation, if you already know what you will do in this situation.[12]

Carlson's starting point here isn't obviously correct. Imagine a situation where I am tempted to steal a piece of candy out of the bulk food section of the grocery store. Suppose I know that I will not do so (say, because I have very strong moral objections to doing so and a very strong will.) This doesn't prevent my moral theory from being action-guiding on this point. So, a theory can be action-guiding even in cases where the agent knows what she will do. But perhaps the trouble is that if NI is false, then in some cases, in order for the theory to *guide* the agent, the agent will have to know what she is going to do, which means that she will not be able to use that information in her deliberations. He will argue that in this case the theory isn't guiding the agent.

So, maybe Carlson's point is that he wants an agent to be able simultaneously to see exactly what the moral theory is instructing him to do, and to be genuinely

[11] I am aware that this involves a fair bit of unrigorous picture-thinking. Incidentally, I think there may be cases where it is true that I do wrong by doing x, where I would have done wrong had I refrained from x, but where it is still true that I *could* have avoided wrongdoing. That is because there may be A-worlds that I can produce that are not the closest A-worlds to me. Worlds where I break the law by jaywalking are closer than worlds where I break the law by committing a great train robbery. It doesn't follow that the latter are inaccessible to me. Consider now a malicious person, who is such that worlds where she swerves to the left to avoid hitting a pedestrian may be more distant than worlds where she swerves to the left to hit a pedestrian, but both are accessible to her.

[12] See *Consequentialism Reconsidered*, p101.

undecided about what to do. But why not accept a different principle that is compatible with a denial of normative invariance: "a reasonable theory should be action-guiding for an agent with complete knowledge of all morally relevant facts in the situation in question except about whether the agent will so act."

Well, Carlson may respond, if whether a particular act is obligatory depends on whether she will actually perform that act, then "whether she will so act" is morally relevant. Technically yes, but suppose she knows the two counterfactuals: if I do A, A is obligatory, and if I do not-A, then not-A is obligatory. In that case, given all the rest of her information, the theory is sufficiently action-guiding. She knows what to do to avoid wrong-doing and to act as she ought. Think about the Huxtable family case above. If they know the two counterfactuals, "If you have the extra child, you will be acting rightly" and "If you do not have the child, you will be acting rightly," then, even if they do not know what they will in fact do, they can proceed with impunity.

For these reasons, the sorts of considerations Carlson mentions seem far from decisive.

But, an objector may complain, "What that boils down to is saying, "Whether you do A or not, you're not doing wrong. Both options are permissible." Isn't that really what is going on here? Why be so weird as to say, "A is wrong, but it wouldn't be wrong if you did it."? What does it mean to say that unperformed A is wrong, unless it means that A *would be* wrong if it were performed? This at the very least seems to suggest that we need a good motivation for denying NI.[13]

Five Challenges to NI

The following are presented as arguments against NI. In each case, the premises are neither obviously true nor universally accepted, but they indicate how much one has to give up in accepting NI.

Inability

Joe is faced with a roulette wheel. He can push a button to set the wheel turning, but how fast it turns and where it ends up is determined by the mechanism. Let's suppose that he will win if the wheel lands on 35 and will lose otherwise. Add that it's a very large wheel with thousands of numbers. Consider first the case where the wheel does not land on 35 and then the case where it does. If it does not land on 35, then Joe could not have made it land on 35. If something important (like his family's financial security) hung on its landing on 35, then, since OIC, it is not true that he ought to have made it land on 35.[14] His inability to get the wheel to land on 35 is the kind of inability that gets one off the "ought" hook. But now imagine that the wheel *had* landed on 35. Many are tempted by the thought that he could, in that case, get

[13] See Krister Bykvist, "Violations of Normative Invariance: some thoughts on shifty oughts," *Theoria*, forthcoming, for discussion of similar arguments. This paper came to my attention after my own paper was mostly completed.

[14] He could have refrained from playing, of course. But suppose that here, unlike in real life, refraining is no better than playing and failing to win.

the wheel to land on 35. Actuality implies possibility, after all. If that's right, then we might say, in that case, he ought to have made the wheel land on 35. So, given that he did x, he ought to have done x; but if he hadn't done x, it wouldn't have been true that he ought to have done x. So, we have a counterexample to NI.

One might object here that, even where the wheel does land on 35, it is not true that Joe can make the wheel land on 35 since it is not properly under his control.[15] Actuality may imply possibility, the thought goes, but it does not imply *ability*. There is no specific point on the dartboard that I can hit from 10 ft away, even though, clearly, if I hit the dartboard, I hit a specific point. I no more have the power to hit that point than any other point. If that is right, then this case loses its force against NI.

But a slight variation on the case does constitute a problem for NI. Suppose the agent is attempting to do something that is in a fuzzy borderline of his ability: e.g., lifting a heavy weight that is neither definitely light enough for him to lift nor definitely too heavy for him too lift. Suppose he tries to lift it, succeeds in lifting it, but in the closest worlds in which he doesn't lift it, (that's because) he cannot lift it. In the world where he fails, he cannot lift it. In the world where he succeeds, it is arguable that he can lift it. The case is importantly unlike the case of the roulette wheel or the dartboard. If there is a moral reason for him to lift it that is strong enough to constitute an ought unless it is cancelled by inability, it seems that we should say that if he does lift it, he ought to lift it; but if he doesn't lift it, it is not true that he ought to lift it. Carlson presents this sort of case as part of an objection to the Principle of Normative Invariance by Wlodek Rabinowicz. Carlson's response to it is as follows:

> The violation of NI in this case is rather special, however, since it concerns an action which would not only lack moral status if it were not performed, but would not even be an alternative for Brown in this situation. Confronted with this type of case, we might weaken NI in the following way:
>
> NI′ If an action is an alternative for P in S whether or not it is performed, then its moral status does not depend on whether or not it is performed."

So, we have what looks like a counterexample to NI, and a revised principle to accommodate it. Let's consider some more purported counterexamples. As we go, let's also consider the question of how *NI′* faces up to these counterexamples.

Chance

Jean-Paul Vessel offers the following case:

> The Demon offers Sam this choice:
>
> If Sam flips the fair coin and the coin lands heads, then the demon will bring about the Good (Pleasure). But if Sam flips and the coin lands tails, then the demon will bring about the Bad (Pain). And finally, if Sam abstains from flipping the coin, then the demon will leave things pretty much the way they are: neither wonderful nor abysmal, rather somewhere in between....

[15] See my "The Rejection of Objective Consequentialism," *Utilias*, 1997.

> When considering whether or not to flip the demon's coin, Sam hopes to perform the alternative that has the better outcome, the alternative that would produce the greater balance of pleasure over pain in the world. So, Sam must decide which of the two alternatives open to him in this situation,
>
> a1 flip the coin
> a2 don't flip the coin,
>
> Has the better outcome.
>
> Suppose Sam doesn't flip.
>
> Was he right?
>
> The answer, however, seems to rest upon the truth values of the counterfactual conditionals utilized by Sam in his moral reasoning about the case, which appear to be the following:
>
> Cf1 If Sam were to flip the coin, then it would come up heads.
> Cf2 If Sam were to flip the coin, then it would come up tails.
>
> If Cf1 is true, then it can be concluded that Sam failed to achieve his utilitarian goals for the truth of Cf1 implies that flipping the coin would have produced the best results possible. But if, on the other hand, Cf2 is true, then Sam has succeeded in performing his utilitarian duty his opting not to flip the coin produces much better results than flipping tails and unleashing the demon's wrath upon our world.[16]

David Lewis's semantics for counterfactuals implies that both counterfactuals are false.[17] According to Lewis, a counterfactual is true if and only if its consequent is true in *all* the closest worlds in which the antecedent is true. Supposing that there can be more than one equally closest world where the antecedent is true, it sometimes happens that some of these worlds are worlds where the consequent is true and some are worlds where the consequent is false. Vessel's case seems like this. There are many ways the coin could be tossed, starting heads down, starting tails down, with this much force or that, and so on. If we imagine these worlds continuing in accordance with the laws of nature, some of them have the coin landing heads and some of them have the coin landing tails.

So, what does that mean about the moral status of Sam's failure to flip? To determine the moral status of Sam's failure to flip we need to compare its results with the results of his alternative action (flipping). But given that it is not true that the results of his alternative action would have been very good, and not true that they would have been very bad (and not true that they would have been indifferent) it seems plausible to say that there are no facts of the matter about what those results

[16] Jean-Paul Vessel, "Counterfactuals for Consequentialists," *Philosophical Studies*, 112 2003 pp. 103–125.

[17] See David Lewis, *Counterfactuals* (Cambridge: Harvard University Press, 1973).

would have been. So, perhaps we should say that there is no fact about whether his failure is wrong, or perhaps we should say that it is not wrong.

Now, suppose instead that Sam (or Pam in Vessel's story) does flip the coin and the coin lands tails. In that case, according to the Lewisian semantics, "If Sam were to flip the coin, then it would come up tails" comes out true, and hence, it was wrong to flip the coin. This is because, in the case in which the counterfactual has a true antecedent, the counterfactual is true if and only if it has a true consequent also. The idea is that the actual world is closer to itself than to any other world. So, the fact that he flipped made it wrong, whereas it would not have been wrong if he hadn't flipped. (At least, perhaps we should say that it would be such that it was not the case that it was wrong, to account for the thought that there may have been no fact of the matter.)

Vessel relies on NI and this case to argue against Lewisian semantics of counterfactuals. But one could instead use this kind of case, plus an insistence that Lewis is correct, to resist NI.

One might object that this argument makes further controversial assumptions, in assuming the truth of objective utilitarianism. But it need not do that. We could vary the case to make it of interest to non-utilitarians. Suppose I have promised to pay two people $50.00 each. Suppose also that I have only half of the funds needed to do so, but someone offers me a bet. If I take the bet and the coin lands heads, I double my money. If I take the bet and the coin lands tails, I lose everything. So,

> If I were to flip the coin and it were to land heads, I would be able to keep both my promises.
>
> If I were to flip the coin and it were to land tails, I would not be able to keep either of my promises.

This case should be compelling to deontologists.[18]

What about Vessel's claim that Lewis is wrong? Are we really entitled to assume Lewis's semantics? Maybe not. But it is widely accepted. It is odd that a moral principle, particularly a somewhat obscure moral principle such as NI, should be able to refute Lewis's semantics for counterfactuals. Shouldn't the plausibility of that theory depend on metaphysical and philosophy of language issues? Shouldn't morality be neutral on that point?

Counterpart Theory

Here's a counterexample to NI.[19]

> I'm faced with a choice between pushing two buttons, red and green. Pushing the red one will produce 100 units of good. Pushing the green one will produce 50.
>
> I push the red one. I do the right thing.

[18] An objector may worry that these two examples assume a sort of objectivism about morality (in the sense of objective utilitarianism as opposed to expected utility utilitarianism.) Below I shall discuss an example that should be compelling both to objectivists and non-objectivists.

[19] Hud Hudson thought of this example and the two that follow.

> However, if I had pushed the green one, it is not the case that the red one would have been right, because in that world there is another option, yellow, which will produce 150.
>
> I have a counterpart who pushes green, but no counterpart who pushes yellow. That counterpart has a counterpart who pushes yellow.

If I push the red button, I ought to push the red button. If I had not pushed the red button, then it would not have been true that I ought to push the red button. Again, this seems to be a difficulty for NI′ as well as NI, as pushing the red button is an alternative in both cases.

This argument rests on counterpart theory, which is very controversial.[20] Counterpart theory is a view (or set of views) that uses the counterpart relation as a replacement for the identity relation between objects in different possible worlds. One's counterpart is a part or inhabitant of other possible worlds whose properties and actions are the truth makers for facts about what one might or would have done. Identity is a reflexive, symmetric and transitive relation. Since the counterpart relation is only a similarity relation, it doesn't have to be transitive or symmetric. This means that A can have a counterpart, B, that has a counterpart, C, although C is not a counterpart of A. Someone could turn this argument on its head and use it to reject counterpart theory. But one could argue that whether counterpart theory is correct should depend, not on a fairly obscure moral principle, but on metaphysical matters.

Altered Past Compatibilism

Here's another counterexample to NI.

> The Principle of Normative Invariance is false if altered past compatibilism is true. Suppose that I give you five dollars because I promised to do so. That was morally required. In the closest world in which I do not give you five dollars that is because I did not promise you the money. In that world giving you five dollars was not morally required (and maybe even was wrong...)
>
> If I give you five dollars, then I ought to give you five dollars. If I had not given you five dollars, it would not have been true that I ought to give you five dollars.

This is a violation of NI. Note that if this objection works against NI, it also works against NI′ as the action of paying you five dollars is an alternative whether I have promised or not.

This argument rests on altered past compatibilism which is very controversial.[21] altered past compatibilism is the view that an agent can act otherwise than she does

[20] For an account of counterpart theory, see David K. Lewis, "Counterpart Theory and Quantified Modal Logic," *Journal of Philosophy* 7 March 68; 65, 113–136. For an account of why it's so controversial, see Trenton Merricks, "The End of Counterpart Theory," *Journal of Philosophy*, 0 03, 100 (10), 521–549.

[21] For some of the debate over altered past Compatibilism, see Jan Narveson, "Compatibilism Defended," *Philosophical Stuides*, July 77, 83–87; John Martin Fischer, "Power over the Past," Pacific Philosophical Quarterly, October 84; 65, 335–350; and Peter Forrest, "Backward Causation in Defense of Free Will," *Mind*, April 85, 94, 210–217.

in spite of the fact that determinism is true, because she can act in such a way that the past would have been different. Someone could turn this argument on its head and use it to reject altered past compatibilism. But one could argue that whether altered past compatibilism is correct should depend on complex metaphysical matters and not on a fairly obscure moral principle.

Minimally Back-tracking Counterfactuals

To avoid the controversial quality of the examples in C and D, let's consider another example:

> Some two hundred of us are ushered into a room to take part in a psychology experiment. We are shown our desks, each of which has a device with three unilluminated buttons on it: Yellow... Red... Green.
>
> The woman conducting the experiment comes into the room and announces the following:
>
> "In just a few minutes, some (and maybe all) of the buttons on your desk will be illuminated. About half of you will see the Red and the Green button illuminated and the other half will see all three buttons illuminated. If a button is illuminated – that means it is live, an unilluminated button is dead... pushing a dead button will do nothing at all. Whether you will be among the first group or the second is simply a matter of chance.
>
> Now, let me first address myself to those of you who will see only two buttons illuminated: In every such case, if you push Red you will cause $100 to be donated to Oxfam, and if you push Green you will cause $50 to be donated to Oxfam. There are no other consequences of any significance to worry about.
>
> And as for those of you who will see all three buttons illuminated: In every such case, if you push Red you will cause $100 to be donated to Oxfam. But here's the fun part... pushing one of the other two buttons will cause $50 to be donated to Oxfam whereas pushing the other one will cause $1,000 to be donated to Oxfam. The catch is you won't know which is which. That will also be determined by chance."
>
> After her speech, I sit anxiously at my desk hoping for a chance to take a risk. That's what I'd do, if all three buttons went live for me. But, unfortunately, I'm at one of the boring desks with just Red and Green as live options, and I push Red (as you would have guessed).
>
> Our objective consequentialist says that I did the right thing. It wouldn't kill him to buy me a beer for it.
>
> But could I have pushed Green? Yes. And the nearest worlds where I do push Green are worlds where chance provided me with three illuminated buttons to choose among. That is, the nearest worlds in which I push Green are not worlds in which my character changes and I try to prevent Oxfam from getting the goods or worlds in which my finger slips or worlds in which I misunderstand the instructions or whatever... they are worlds where chance provided me with

> the opportunity to take the risk of getting $50 less than I knew I could get in the hopes of lucking out and getting $900 more than I knew I could get.
>
> Our objective consequentialist says that in the nearest world where I push Green, it would have been wrong to push Red (the action I performed permissibly in the actual world). It doesn't matter whether or not I guessed correctly that Green was the $1,000 button, either – it's enough that the best button was other than Red.[22]

So, it seems I did the right thing by pushing red. If I had not pushed red, doing so would have been right also.

Objection: This example, like the previous two, seems to presuppose some weird metaphysical phenomenon such as counterpart theory or altered past compatibilism.

Response: no, it does not. It relies only on the truth of some minimally backtracking counterfactuals. In the story about the buttons, suppose it is decided by an indeterministic mechanism whether two or three of the buttons on my desk are illuminated. Suppose it is decided just a fraction of a second before I am faced with my choice. It is plausible to suppose that the smallest change from the actual world to a world (set of worlds) where I do not is one where the mechanism indeterministically illuminates the third button.

It may be objected that some form of compatibilism is being assumed here, since incompatibilism implies that for an agent to act freely (and hence, for an agent to perform an act with moral significance) she must have the ability to produce a world exactly like the actual world up to the time of action.

Although I do endorse incompatibilism about freedom and moral responsibility, I would like to make two points by way of response here. First, even if the agent of act P has the ability to actualize a certain world exactly like the actual world up to the moment of action but in which she does not do P, it doesn't follow that this is the closest world in which she fails to do P. In the example given above, the agent may have had the power to refrain from pushing the red button even when faced with only red and green buttons, but it may also be true that a closer 'non-red-button-world' is one where he has three options instead of two. Secondly, I think that one can have enough freedom for moral responsibility with respect to some choice if one is, in that very moment, determined to act a certain way, as long as the determination has been made by earlier free choices one has made.[23] So, for example, suppose our agent has developed his character by choices made in genuine indeterministic choice situations so that he is now firmly disposed to push the most charitable button; or suppose he has simply made up his mind in the minute or so he had while waiting for the buttons to be illuminated. It seems that this still counts as a morally significant choice for which the agent can be held responsible.

In conversation, Ryan Wasserman objected to this last example like this: Backtracking counterfactuals are usually false, or at least, depend on weird contexts. In

[22] This is the promised example that should appeal both to those who accept, say, objective consequentialism, and those who accept expected utility consequentialism. In the case where the agent chooses to take the risk of pushing the button that has a 0.5 chance of causing $1,000 to be sent to OXFAM, this seems to be the right choice from the point of view of expected utility, and the right choice if, unbeknownst to the agent, the button he pushes will in fact lead to the $1,000 being sent to OXFAM.

[23] For more on this idea, see Peter van Inwagen, "When is the Will Free?" *Nous Supplement: Philosophical Perspectives*, 1989; 3.

support of this, he cites David Lewis who writes: "Backtracking counterfactuals, used in a context that favors their truth, are marked by a syntactic peculiarity. They are the ones in which the usual subjunctive conditional constructions are readily replaced by more complicated constructions: "If it were that.. then it would have to be that..." or the like.[24]

Lewis does allow that in unusual circumstances, or with some manipulation of the context, we might agree to accept some backtracking counterfactuals. Here's an example that seems not too weird. "If I were in California a second from now, I would have recently taken a plane trip," seems more plausible than, ""If I were in California a second from now, I would have traveled faster than the speed of light." The latter holds the past fixed, while the former involves some backtracking. Surely the latter is more plausible?

A second move Wasserman made was to argue that there is a kind of incoherence in the example in that, on the one hand, in the original case, the agent congratulates himself on doing the right thing—he pushed the red light, giving $100 to Oxfam, when his alternative was to give only $50. In other words, part of what makes it the case that he did the right thing was that if he had done otherwise, he would, at best, have given only $50. However, in subsequent telling of the story, we are told that this is not the case. It is not true that if he were to have pushed green, he would have given only $50. Since, if he had pushed green, he might have given $1,000.

In response, we might note that counterfactuals are notoriously malleable. What I need to argue to make the example work for me is that the agent pushed the red button because pushing the green button *given that only two buttons were illuminated* would lead to less good. But, at the same time, if he had pushed the green button, more good would have been produced. Clearly, to make this work, we need two different antecedents: "If I pushed the green button, and only two buttons were illuminated..." and "If I pushed the green button..." on the assumption that this second antecedent takes us to a world where all three buttons are illuminated. Are we entitled to these two different antecedents? I have already argued that the second is the correct one with respect to the question of what would have been the case if I had pushed the green one.

I need to argue that the first one is the correct one to use in assessing the moral status of my action of pushing the red one in the actual circumstances. I believe it is because the number of buttons that is illuminated is not something the agent is responsible for or something that he has any power to change. He has to act in light of how that fact actually stands independently of him. But it is consistent to say that the closest world where he pushes the green button is one where this fact is different, rather than one where he makes an immoral choice.

Variations on a Theme

Just as Carlson responded to the first counterexample by shifting to a variant on NI, namely NI′, perhaps we could tweak NI or NI′ to avoid the other counterexamples. Recall his initial principle and his variant:

NI: An action's moral status does not depend on whether or not it is performed.

[24] In Lewis "Counterfactual Dependence and Time's Arrow," *Philosophical Papers II* (Oxford U. Press, New York: 1996) p. 34.

And

NI′: If an action is an alternative for P in S whether or not it is performed, then its moral status does not depend on whether or not it is performed.

In response to Vessel's case about Sam and the coin and the demon, a defender of normative invariance could offer:

NI″ If an action is an alternative for P in S whether or not it is performed, and if the action *will have determinate results whether or not it is performed*, then its moral status does not depend on whether or not it is performed.

In response to counterexamples C and E,

NI‴: If an action is an alternative for P in S whether or not it is performed, and if the action will have determinate results whether or not it is performed, *and if the action has the same alternatives*, whether or not it is performed, then its moral status does not depend on whether or not it is performed.

In response to the counterexample about altered past Compatibilism,

NI⁗ If an action is an alternative for P in S whether or not it is performed, and if the action will have determinate results whether or not it is performed, and if the action has the same alternatives, whether or not it is performed, *and if the reasons for and against the action are the same whether or not it is performed*, then its moral status does not depend on whether or not it is performed.

This principle is starting to get a little bulky. Perhaps it could be abbreviated to read:

NI*: If an action is the same in all morally significant respects whether or not it is performed, then its moral status does not depend on whether or not it is performed.

Now (I think) we have a truth. Any purported counterexample would be subject to the objection that the feature which explained why x was right in the actual world and wrong in some other world would be a morally significant difference between the worlds.

This principle seems true, but how useful is it? Think back to one of the cases we started with. Suppose the Huxtable family is trying to decide whether to have another child. If they do, they reason correctly, the child will be happy, but each member of the family will be a little bit worse off. We might even suppose (if this is coherent) that the total amount of happiness or other good the new child will enjoy will be greater than the sum of what is lost by the other family members. They decide not to have the child. According to the actuality principle, they did the right thing. The comparison between the happiness of the actual family members in the two scenarios is what matters morally. The hypothetical happiness of the non-actual child is irrelevant.

But now change the story. Suppose that the Huxtables decided differently. They decided to have a child and did so. Now, this child, call her Rudy, is actual. Her happiness is sufficiently great to outweigh the sum of all the small losses in well-being each of the family members sustained. Now, her well-being does count.

Here we are comparing two worlds, one with, and one without Rudy. The Actuality principle implies that the action of producing Rudy is right (or permissible) in the world where they do so; but not right or permissible where they do not. What does NI* imply about this case, and hence, about the actuality principle? Well, that hangs on the question, “Is Rudy’s actuality a morally significant difference?” or “Does the actuality of an extra person make a morally significant difference between the worlds?”

That’s a difficult question, but note two things. First, it is very tempting to say yes. And second, the principle of Normative Invariance, as it has been modified, is unable to help us answer it.

Conclusion

The Principle of Normative Invariance, attractive as it is at first glance, is subject to a number of powerful counterexamples. Attempts to modify the principle to avoid these counterexamples deprive it of any interest or power. It seems, therefore, that philosophers should not be so quick to invoke the principle. This is important news, because it undercuts an argument regarding our obligations to future generations that has been thought to be decisive. In particular, the main objection to certain actualist views about our obligations to future generations derived from the Principle of Normative Invariance.[25]

If this principle is undercut, these views remain viable options.

Acknowledgment Thanks to Ryan Wasserman, Ty Barnes, Ned Markosian, two anonymous referees for Philosophia, and especially, Hud Hudson for all the help. I am also grateful to Western Washington University for providing me with professional leave which enabled me to complete this project.

[25] See Parsons, Arrhenius, Narveson, *op. cit.*