Doing Our "Best"? *Utilitarianism, Rationality and the Altruist's Dilemma*

1. Introduction

Suppose you think – as I think – that what matters centrally in ethics is making the world a better place. Well, when I say *making* the world a better place, I don't think it matters that *I* make the world a better place – that would be a terribly self-important attitude. I just want the world to *become* a better place. When I talk of the world being a good place, I'm thinking of it being good *for* its inhabitants – benefitting them, giving them welfare, perhaps making them happy. But the details of this axiological question don't matter for what I want to say today.

It is easy to suppose that, if we want to make the world a better place, we each ought to *do our best* – perform the actions, out of those open to us, with the best expected outcomes. In other words, it might seem that we should rationally become *Act-Utilitarians.* But whereas the notion of goodness is an absolute one, the notion of a *best* option is relative – it is defined relative to a set of options. And determining what count as the options open to a rational agent is a far from simple matter. After all, what *my* options – as a rational, moral agent – are, depends, in many cases, on what *other people* are going to do. But many of *those* people are *also* rational moral agents. So what *they* think they rationally ought to do will, in part, determine what they *will* do, and so constrains what options *I* have. And, one hopes, in many cases what they *think* they have most reason to do will reflect what they *in fact* have most reason to do.

Thus rational moral guidance cannot just be a matter of *reacting* to the world as we find it, as Act-Utilitarianism supposes – for moral reasons partly *shape* the world to which we are reacting. And there are ways that moral agents could collectively construe their options which diverge from the Act-Utilitarian framing, but which make *better* outcomes available than if they had been strict Act-Utilitarians. So, even if our only premises are fundamentally Utilitarian ones, we should reject Act-Utilitarianism as a complete account of moral reasons. And unlike historical "Utilitarian" objections to Act-Utilitarianism, there is no way, I argue, of responding to *this* argument by saying that other principles are merely "rules of thumb", which guide us to do the acts objectively recommended by a secret, "Esoteric" Act-Utilitarian morality. My argument is that Act-Utilitarian reasons are not, uniquely, *objectively* optimific.

The case I'm going to focus on concerns the morality and rationality of political action. I pick this case for a reason. For, although we now tend to teach Utilitarianism as a view in moral philosophy, this presents a narrow view of Utilitarianism in the history of practical thought. While the early Utilitarians did articulate versions – criticised by later philosophers for their seeming ambiguity – of the "Principle of Utility", they were not primarily concerning with personal morality. John Stuart Mill, Harriet Taylor, James Mill, Henry Godwin, even – despite his association with "Act-Utilitarianism" – Jeremy Bentham, were primarily interested in questions of social, legal and political reform. It was to such structural questions that they devoted most of their efforts and writings, seeking radical systemic reforms of society: they worked for such goals as the emancipation of women, the expansion of the franchise, the protection of free speech, the disestablishment of the Church, even, in Godwin's case, the abolition the state, and, in Bentham's private notebooks, the legalisation of homosexuality. To be sure, some of their political ideas were downright weird or even objectionable, and the boundary between moral and political philosophy is hard to define sharply. Nonetheless, it seems fair to say that, for the early Utilitarians, questions of personal moral decision-making played second fiddle to visions of social and political reform.

By contrast, a tradition running from the work of Henry Sidgwick, through contemporary philosophers such as Derek Parfit and Peter Singer, to the Effective Altruism movement of today, reverses that emphasis. This approach sees Utilitarianism as, primarily, a moral theory of how individuals ought to behave. On this view, political questions naturally become questions of applied personal morality – most often, questions about the best way for individual Act-Utilitarians to proceed in engaging with the political realm. There has long been a quandary about how "secondary" principles (such as those defending liberty), which the older Utilitarians used to guide social and political reform, are supposed to relate to the "Principle of Utility". But the tension between Act-Utilitarianism, as a personal morality, and Utilitarianism as an approach to radical political reform, has been made clearest by the contemporary Effective Altruist movement. For, as they have argued, it is just not clear at *all* that a rational Utilitarian altruist should concern herself with the pursuit of political reform. Given the reliance of political change on the cooperation of many others, moral rationality should guide the altruist to focus on individual charitable giving, rather than gambling her efforts on something so dicey as the project of changing society.

This surprising thought is my starting point. I want to suggest that history has been too harsh on the early Utilitarians for their failure to articulate a precise account of the Act-Utilitarian account of reasons. For the understanding of Utilitarian moral reasons that I want to advance is one in which moral reasons just *are* ambiguous, and in which questions of personal morality cannot – even at a theoretical level – be decoupled from issues of political and social reform.

But first, I want to tell you a story.

2. The Altruist's Dilemma

Henry wanted, more than anything else, to make the world a better place. He aimed to use rationality and evidence to identify the most effective use of his time, energy, and considerable skills. Not for him a sentimental attachment to lost causes or symbolic protests. Henry, in his love of mankind, wanted to make a difference.

On leaving University, Henry's admirable commitments led him, like many of his most serious and intelligent peers, to seek advice from the leading lights of *Effective Altruism*. But there he found advice that surprised him. He had always assumed that the forum for those who want to change the world for the better was the political realm – perhaps not electoral politics itself, but, at the least, fighting and advocating for the structural reforms that society, he was sure, so desperately needed in order to arise from the morass of preventable human suffering into which it was fallen. Of course, political and economic questions are contentious, and Henry was wise to this fact; nevertheless, there were reforms – concerning climate change, access to healthcare and education, and foreign aid – which seemed, on careful analysis of the best evidence he could find, clearly such as to benefit people, were they to be enacted.

But there, he discovered, was the catch. The organisation to which he had turned for guidance, *Doing the Best We Can* (DBWC), agreed with him that these reforms *would* be exceptionally valuable, *if* they could be brought about. But, they pointed out, to achieve such reform would take far more than the efforts of any one Henry, no matter how intelligent and committed he should be. It would take the cooperation and commitment of a great many people for any of the campaigns Henry dreamt of to have the slightest chance of succeeding. And this, sadly, was – or so the evidence suggested – unlikely. Given this unlikelihood, Henry concluded, regretfully, that seeking political change was not, after all, his best option in his search to change the world for the better. DBWC advised him that he would do far better to seek a conventional job, make plenty of money, and donate it to highly effective charitable organisations. Given the power of their argument, this seemed to Henry to be the only reasonable conclusion.

Or, at least, so he thought at first. But then something struck him. After all, *a great many of Henry's peers from university had also turned to DBWC for guidance.* These were morally motivated, intelligent people just like Henry, who would have pursued political reform if they had thought that this was the path recommended by rational morality. In convincing them to eschew politics in favour of lucrative employment and charitable giving, DBWC had *reduced* the number of potential co-operators Henry might have found in his political ventures, and thus made it *even more unlikely* that any such political campaigns might succeed. After all, unlike some avenues for improving the world that offered a low probability of a great reward – things like nuclear fusion research – there was nothing *intrinsic* to political change that made it improbable. The probability of success was almost entirely a function of the willingness of agents to pursue it.

Thus DBWC's pessimism about political change, which once struck Henry as impeccably rational, now appeared as a dangerously self-fulfilling prophecy. DBWC had encouraged Henry and his peers to think rationally about how best to use their resources to change the world for the better. But the effect of their following this advice, he worried, was to change the world for the worse. On the other hand, if DBWC had *encouraged* Henry and his peers to pursue politics, then, given their numbers and collective talents, the probability of effecting meaningful change would have been quite high. Given the immense value of such change, this would, Henry judged, have been a better result.

At first, Henry suspected that the problem lay in the role of DBWC. Even if the best option for Henry and each of his peers was to follow the pessimistic path to apolitical altruism, perhaps, he reasoned, DBWC should have thought more carefully about its *own* causal effects. Knowing that it might influence a great number of highly motivated and intelligent people should have informed the advice that DBWC decided to give, and so perhaps it really ought to have told Henry and his peers to pursue politics, even if that was not the thing that any of them really had most reason to do. In a sense, DBWC should have hidden the truth about moral reasons from its audience in service of the greater good.

But this, he concluded, was a red herring. After all, *the only thing DBWC had done was to tell each of them what they had most reason to do*; it had no power beyond the power of rational advice. And neither Henry nor his peers would have followed that advice had they not *agreed* that it was rational. In other words, any of them might – perhaps, should! – have reached the same conclusions on their own. And if it would have been best for DBWC to try to *get* each of them to pursue politics despite the poor individual odds of success, this was also something that they might have worked out by themselves – as, indeed, Henry was doing. And yet this only increased Henry's perplexity. How could it be that the reasons that it would be best for all of them to follow were not reasons for each of them to do what was best?

Henry had always been motivated by the simple ideals of rational altruism – and he had always supposed them to be just that: *simple.* And yet now he faced the paralysing conclusion that rational altruism might offer him no clear guidance at all. On the one hand, it did seem rational to think as DBWC had recommended, and eschew politics in favour of the more reliable option of charitable giving, with its higher expected utility. But if it was rational for Henry to think and act that way, then it was equally rational for others to do the same, which would make political reform even more unlikely – and that would be a bad thing!

On the other hand, reflecting on *this* fact made it look rational to think in a more collectivist manner, and pursue political reform despite the individually poor odds of success. But, of course, *that* would only be a good thing if other people *did* in fact follow this reasoning – and Henry had no way of knowing that they would do this. His peers from university were rational altruists just like him. If moral rationality recommended just *one* option, then that would be some basis for predicting how the others would behave, in turn helping Henry settle on what to do. But now Henry saw that there were two, fundamentally divergent, forms of moral reason that he and his peers might follow, and the only thing determining which would be *better* to follow was the fact of what the other were *going* to follow. It would be best for any of them to pursue politics if *the others* were also going to pursue politics, and it would be best to *treat this fact as a reason* to pursue politics if and only if *others* were also going to treat this fact as a reason.

Henry's dilemma ran deep. It wasn't just that he wasn't sure which option to pick in this situation. He no longer felt sure what it meant to be a rationally altruistic in the first place. He wanted to do his best. But what *was* his best option?

3. Who Faces the Altruist's Dilemma?

Henry's thinks that what matters most, in moral terms, is how good, or happy, the world is. And, moreover, he thinks that the moral reasons that apply to him and other moral agents derive from this – what they each have reason to do, morally speaking, is fixed by the ultimate moral goal of promoting the good.

There are two core thoughts here:

> *Axiological Utilitarianism:* What matters most, morally, is welfare; welfare is what is good, and the more welfare people experience, the better.

> *Meta-Deontic Utilitarianism:* There are moral reasons which apply to agents, and the contents of these reasons is fixed by facts about promoting welfare.

I think that these principles are compelling. Indeed, even people who don't think this is the *entirety* of morality might agree that there is a significant portion of moral life concerned with making the world a better, happier place, in the most effective manner possible. These people should feel the pinch of Henry's dilemma too. Even if there are side constraints – concerning rights, justice and so on – that delimit ways in which we can permissibly pursue the good, it seems unlikely that these will settle the question of whether we should prefer the reliably beneficial option of individual charitable giving over the seemingly more uncertain approach of seeking political change.

However, the two principles just stated do not, all by themselves, generate the dilemma. For that, I want to tell a story of how the basic commitment of Axiological Utilitarianism has led philosophers towards the Act-Utilitarian principles espoused by DBWC.

4. And You May Ask Yourself – How Did I Get Here?

I think most of those who identify as Utilitarians will accept *Axiological Utilitarianism*. Moreover, many philosophers not strictly classified as Utilitarians seem to accept something like this too. For example, "proto-" Utilitarians such as Hume seem to think that what ultimately matters in morality is utility (even if he wouldn't accept the hedonist reduction of utility to happiness).

However, by itself, *Axiological Utilitarianism* doesn't to help us decide what to do – how to live our lives, how to behave, how to make morally rational choices. Indeed, this is precisely the fault that many find with Hume's moral philosophy – it tells us that virtues are those traits of character that are useful or agreeable to self or others – that's to say, which promote utility either directly or indirectly – but it doesn't directly give us action-guiding reasons. But wouldn't a morality that gives practical guidance be more *useful* than one which does not? In other words, many people think that morality must have a *deontic component* – it must offer an account of reasons for action which settles which choices moral agents ought to make. This thought leads us to the second principle I mentioned: *Meta-Deontic Utilitarianism.*

However, even *Meta-Deontic Utilitarianism* doesn't, itself, tell us what moral reasons we have. It just says that, *whatever* reasons we have, these are *somehow* derivative of the goal of promoting welfare. But that "somehow" is hard to interpret. Here is a plausible interpretation of the content of the reasons that Meta-Deontic Utilitarianism ought to recommend:

> *Optimific Reasons:* Morality assigns to agents the reasons that it is best for them to follow; the optimific moral reasons are the reasons that will guide agents in such a way that leads to the best outcomes.

In other words, we have the moral reasons which it is best – in the sense of promoting happiness or welfare – for us to have.

What reasons are these? The early Utilitarians often formulated the "Principle of Utility", in ways that seem ambiguous between Act-Utilitarianism and "Indirect" Utilitarian views, such as Rule-Utilitarianism. It's common to understand Indirect Utilitarianisms as theories of the action-guiding moral reasons that apply to each agent. So here's a construal of these views, construing each as theories of reasons for action:

> *Act-Utilitarianism*: Each agent has reasons to choose the act, out of those open to her considered individually, that has the highest (expected) utility.

> *Indirect-Utilitarianism*: Each agent has reasons to perform acts selected by rules/principles/virtues/motives that (would) lead to the best (expected) outcomes if (all/most) agents (obeyed them/tried to obey them/promulgated them).

Obviously, the construal of Indirect-Utilitarianism given here is ambiguous on multiple counts. But that doesn't matter for our purposes. Because Act-Utilitarians (following Lyons 1965) have posed a straightforward dilemma to defenders of any form of Indirect-Utilitarianism.

> *With Us or Against Us?* Either the recommendations of Indirect-Utilitarianism are identical to the recommendations of Act-Utilitarianism, or they are not.

In cases where some apparent alternative to Act-Utilitarianism gives the same advice, Act-Utilitarians can dismiss the disagreement as chimerical. In cases where they diverge, they can appeal to the following principle, which seems to follow from *Optimific Reasons*:

> *No Rule Worship*: If any account of reasons for action tells us to choose the options with (predictably) worse outcomes than some alternative account of reasons for action, then we should reject this account of reasons for action.

If we act in a way that leads to predictable net harm to others, *just because* some rules, principles or standards of virtue seem to recommend this, then it seems that we have put these rules above the goals that they were supposed to serve. And that seems incompatible with *Optimific Reasons*, and *Meta-Deontic Utilitarianism*.

Some Rule-Utilitarians, such as Hooker (eg Hooker 2000), argue that their theory better coheres with our moral intuitions or common sense; but, to that extent, their motivation is *not* the sparse one articulated so far. Though such theories may make reference to the promotion of utility in determining which rules or norms they accept, their ultimate motivation is *not* just to find whatever account of moral reasons is optimific. Rather, the desire to respect common sense serves as an independent normative goal. Thus, these theories are not *Meta-Deontically* Utilitarian – they do not fully subordinate reasons to the ultimate goal of promoting the good.

The *Rule Worship* charge assumes that Act-Utilitarian reasons *are* the *Optimific Reasons*. We can call this assumption the:

> *Equivalence Thesis:* The reasons that are best for agents to follow are, uniquely, reasons for each agent to select (and act upon) the option with the best (expected) consequences, out of those open to her. The *Optimific Reasons* just are *Act-Utilitarian Reasons*.

If the *Equivalence Thesis* is true, then the charge of *Rule Worship* against non-Act-Utilitarians stands. If Indirect-Utilitarianism is "Against Us" – if it posits an account of reasons that diverges from that of Act-Utilitarianism, then Indirect-Utilitarians are Rule Worshippers, since their account of reasons is not optimific, and hence violates *Meta-Deontic Utilitarianism.*

Now we can see how Henry's commitments led him to his dilemma. He accepts *Axiological Utilitarianism*, and *Meta-Deontic Utilitarianism,* leading to the conclusion that he should follow *Optimific Reasons*. And Henry initially accepted the *Equivalence Thesis* – he thought that these reasons must be Act-Utilitarian in form. And those reasons seemed to lead him to the conclusion that he ought to eschew politics in favour of some more reliable means of promoting the good.

5. Objective Reasons and Hidden Principles

Henry's predicament should lead us to question the *Equivalence Thesis*. Henry and the other followers of DBWC are rational, intelligent Act-Utilitarians: and yet, if they *all* follow their Act-Utilitarian Reasons, then they will – or so it seems – each eschew politics in favour of individual charitable giving. But had they been guided by different principles, perhaps they would have pursued politics instead. And as Henry recognised, the *very* best outcomes would be brought about only by politics, not by charitable giving.

We might, then, be tempted by the following two thoughts:

1) In *The Altruist's Dilemma*, the Optimific Reasons are reasons to pursue politics.

Furthermore, if we agree with DBWC's original advice to Henry, we will also accept that:

2) In *The Altruist's Dilemma,* Act-Utilitarian Reasons are not reasons to pursue politics.

If we grant 1) and 2), then it looks as though we should conclude that the *Equivalence Thesis* is false, and thus that the argument from the *Equivalence Thesis* to Act-Utilitarianism can be denied.

Act-Utilitarians may protest that this argument goes too fast. After all, there have been many arguments that agents *attempting* to employ the Act-Utilitarian account of reasons as a guide to deliberation may, in practice, predictably produce sub-optimal results. However, as they argue, this does not imply that the Act-Utilitarian account of reasons is false. They appeal to Railton's (1984) distinction between *Objective* and *Subjective* Utilitarianism. We can thus distinguish two senses in which some view of reasons might be optimific:

> *Optimific Objective Reasons*: A set of moral reasons is optimific if the (expectedly) best results are brought about if agents *actually do what they recommend*.
>
> *Optimific Subjective Reasons:* A set of moral reasons is optimific if the (expectedly) best results are brought about if agents *attempt to employ them as a guide to action.*

Act-Utilitarians say that theirs is a theory of *objective* reasons, and that the Equivalence thesis is true as a claim about optimific objective reasons. It does not impugn a theory of objective reasons if irrational, imperfect agents *fail* to do what it really recommends when *attempting* to employ it in deliberation. If it is in fact true that Act-Utilitarian reasons are not optimific taken as *subjective* reasons, then agents should employ some other decision-making procedure.

In light of this, Act-Utilitarians can also appeal to the idea, derived from Sidgwick and developed by Singer and Lazari-Radek (2014), that theirs might be an *Esoteric Morality* – people might act better if they did not believe in Act-Utilitarianism at all. In that case, convinced Act-Utilitarians needn't try to persuade others to *become* Act-Utilitarians themselves – they should merely give other people whatever moral advice would likely bring about the best outcomes.

Putting these ideas together, Act-Utilitarians might accept both 1) and 2), but argue that DBWC should have *told* its audience that what they had most reason to do was to pursue political reform. They could argue as follows:

i) The expectedly-best outcome would be brought about if all of DBWC's audience were to pursue political reform.

ii) If DBWC were to tell its audience that they each had most reason to pursue political reform, then they would indeed each pursue political reform.

iii) The option open to [whoever is in charge of] DBWC with the highest expected utility is to *tell DBWC's audience to pursue political reform.*

Since it is not morally important that any agent subjectively *deliberate* in Act-Utilitarian terms, then, if the best outcome is one where everyone pursues politics, Act-Utilitarians reasons recommend that *DBWC* should give whatever advice will bring this about – even if that advice does not state the reasons that Act-Utilitarianism would give to DBWC's *audience*. So *The Altruist's Dilemma* does not threaten Act-Utilitarianism.

This might make sense if DBWC's audience were blindly obedient followers – if they had no capacity to evaluate claims about reasons for themselves. But, as I have told the story, this is not the case – ii) is false. Let us assume that the opposite is true – that DBWC's audience are ideally rational, altruistically-motivated agents. As Henry observed, he and his peers would only follow the advice *if they could see that it was rational.* So it cannot be an option for DBWC to solve the dilemma just by *making* followers do the best thing. Any advice DBWC gives to its followers is advice they could have worked out for themselves; if DBWC's advice moves them, it is only because they can see that it accurately states the reasons that they have.

Indeed, the stipulation that Henry and his peers are ideally rational altruists undercuts the move to Esoteric manipulation in the first place. It makes sense for Act-Utilitarians to give advice *other* than statements of Act-Utilitarian reasons in cases where the recipients of the advice are likely to make errors of calculation, succumb to temptation, or emotionally reject Utilitarian reasoning. But none of these conditions apply here. Likewise, Railton's distinction between Objective and Subjective Reasons appeals to the harm that an excessively impartial mindset can have to human relationships. But no such issues are relevant to *The Altruist's Dilemma*.

Indeed, nothing in *The Altruist's Dilemma* rides on the ways that Henry and his peers deliberate subjectively – what matters is what actions they choose. If 1) is true, then the correct *action* for them to choose is to pursue political reform, but if 2) is true, then Act-Utilitarianism does not attribute to them reasons to pursue political reform. It may be consistent with Act-Utilitarianism to say that we should hide the truth of Act-Utilitarianism from people in order to get them to do the actions that Act-Utilitarian reasons objectively recommend; but could it be consistent with Act-Utilitarianism to say that we should manipulate others in order to get them to do something *other* than what Act-Utilitarian reasons objectively recommend?

7. The Agent-Neutrality of Utilitarian Reasons

To see what has gone wrong here, we can appeal to the following principle. Since Act-Utilitarian reasons are supposed to be agent-neutral, it seems that we should accept:

> *Reasons Transmission:* If A has most objective moral reason to *bring it about that* B performs P, then (other things being equal[1]) B already has most objective moral reason to P.

In other words, if DBWC has most reason to *get* its followers to pursue politics, then they have most reason to pursue politics. Here is a way of seeing why this principle must be true:

> *No Immoral Morality:* If the objectively morally best thing is that I do P, it cannot be that I act objectively immorally in doing P.

If we think of rationality in terms of instantiating particular patterns of decision-making, then familiar examples such as Parfit's (1984) *Robber* scenario show that there can be cases of *Rational Irrationality* – cases it is rational for an agent to make herself think irrationally. But given the Utilitarian commitment agent-neutrality, there cannot likewise be cases of moral immorality – if it is objectively moral for me to make myself or someone else act in a particular way, then it is objectively moral for them to act this way.

Here is an important exception to the *Transmission* principle:

> *Buridanic Coordination*: In cases where agents B and C need to coordinate on some action, and both do *either* P *or* Q, such that the result of B and C both doing P is just as good as the result of B and C both doing Q, whereas no good results if they each choose different options, then A's advice can make a difference to what B and C have most reason to do, compared to what they previously had reason to do.

In a Buridanic case, B and C each have sufficient reason to do either P or Q, but no decisive reason to choose between the two. Likewise, A has sufficient reason to advise both B and C to

[1] There might be extraneous factors complicating this inference – for example, if there is some independent benefit caused by A's act of trying to influence B, or if some third party will punish B if she performs P without being compelled. These *ceteris paribus* riders are not relevant for the argument I'm making.

do either P or Q, but no decisive reason to choose between the two. But they all have decisive reason to bring it about that B and C coordinate on one option or the other. If B and C cannot do this on their own (for example, if they cannot communicate), then A can break the tie for them. However, in this case, A's advice changes the normative landscape only by providing B and C with an extra piece of *information*: A makes one of the options *salient,* in Schelling's (1960) sense:

> *Rational Salience*: If an agent must perform either P or Q, and has equally strong reasons to perform both P and Q, she may rationally pick whichever is more salient.

If A tells B and C to do P, then they should do P, not because it is intrinsically a better option, but because a good outcome requires coordination, and they now each have good evidence that the other will plump for P over Q, since P is now mutually salient: each knows that the other is rational, and that rational people will use salience to break Buridanic ties.

But this principle isn't relevant to *The Altruist's Dilemma*. If 2) is true, and if Act-Utilitarianism is true, then Henry and his peers do not have equally strong to pursue politics and to devote themselves to charitable giving. Rather, they each have determinately strongest reason to pursue charitable giving. So they cannot rationally use salience to break the tie between the two, since there is no tie in the first place. Given this, plus mutual knowledge of rationality, none of them can expect the others to pursue politics just because DBWC tells them to, and so none of them has reasons to pursue politics in response.

Contrariwise, if DBWC were to give Henry and his peers advice, it would not be a matter of *Buridanic Coordination*. DBWC shouldn't just toss a coin to decide which choice to direct its audience to. Rather, since political reform is objectively better than collective charitable giving, DBWC should *determinately* prefer to advise its followers to pursue reform. But if DBWC has determinate reason to give this advice, then, given *Reasons Transmission,* it must have been true that they already had determinate reason to pursue political reform anyway.

So we can see why Henry's first thought about DBWC's role was, as he later realised, incorrect. If it is true that the best outcome is for him and his peers to pursue politics, then the problem for Act-Utilitarianism cannot be solved by appeal to the causal, coordinating role of DBWC. Thus, we cannot appeal to *Esotericism* or *Buridanic Coordination* to counter the argument that, if 1) and 2) are true, then Optimific Reasons are *not* Act-Utilitarian Reasons.

8. What Does Act-Utilitarianism really advise?

The Altruist's Dilemma is inspired by the "Institutional Critique", advanced by Srinavasan (2015) and others, which charges that Effective Altruists have, in their prioritisation of charitable giving, wrongly ignored possibilities for valuable systemic change. Effective Altruists (eg Berkey 2018) have often responded that this is unfair. If seeking systemic change isn't the most effective way to make the world a better, then they don't see why they should endorse it. But if it *is*, then surely it already follows from their principles that they should endorse it, and so the Institutional Critique is not a critique of Effective Altruism or its Act-Utilitarian underpinnings, but merely of the misapplication of their principles. In other words, either 1) or 2) must be false – either the Optimific Reasons *are not* reasons to pursue political reform, or Act-Utilitarian reasons, properly understood, recommend the pursuit of politics. We turn now to these possibilities.

Maybe 2) is false. Perhaps DBWC was wrong in its initial assessment, and Act-Utilitarian Reasons really do direct each of Henry and his peers to pursue politics. We have already stipulated that political reform is the outcome with the greatest actual utility, if it can be brought about. And if enough people choose to pursue it, then it can indeed be brought about. So the question of whether it is the option with the greatest *expected* utility depends on what Henry can rationally expect others to do.

DBWC's advice hinged on the assumption that it was unlikely that enough other people would join in political struggle. But perhaps *that* assumption is the problem. Perhaps DBWC had not accurately assessed how many followers it had, or underestimated their altruism and moral rationality. Perhaps, in the situation as we have *now* described it, where Henry (and DBWC) know that there are sufficient other moral, rational altruists out there to give political reform a good enough chance to be worth it *if only they decide to pursue it*, then political reform *is* the option recommended by Act-Utilitarian reasons.

Of course, if pursuing political reform was the thing the others had most reason to do, then Henry would expect them to do it, and so would pursue reform in himself. But that would beg the question. Just as Henry is trying to work out what *he* has most reason to do, so too are the others. We must ask *for each of them* whether pursuing politics is the option with the best expected utility, without having *already* determined that the others have reason to do the same. In other words, our question is – what is the option with the best expected utility *if the relevant agents are all rational Act-Utilitarians, and know this about one another?*

To answer this question, we can turn to work on the *Hi-Lo Game*, described by game-theorists and economists such as Bacharach (2006), Gold & Sugden (2007), Colman and Gold (2020) and Sugden (2015), and theorists of Utilitarianism, including Gibbard (1965), Regan (1980) and Woodard (2017). In the Hi-Lo game, Agents A and B must select between Actions 1 and 2, with the following payoff matrix, with payoffs defined as arbitrary units of impartial utility:

HI-LO (moral)	A chooses 1	A chooses 2

B chooses 1	20 units of utility	0 units of utility
B chooses 2	0 units of utility	10 units of utility

The value of the options open to each agent depends on what the other agents choose. There are multiple "Nash Equilibria" – outcomes where no agent can improve the situation by changing what she does, given what the others do. But one Nash Equilibrium is better than the others (it is "payoff-dominant"). It is for the best if the agents converge upon this optimal equilibrium (the "Hi" outcome). But can each of the agents expect the others do to this? If they are rational maximisers, it is not clear that they can. If each agent plays her part in bringing about the suboptimal Nash Equilibrium ("Lo"), they have each done the best act open to them individually. Since each agent knows that all the others can see this, none of them can rationally expect that the other agents *will* do their part in collectively enacting "Hi". And so, picking Option 1 is not determinately rational.

Now, you might think that the problem here arises from the assumption that each agent uses as her sole criterion of choice the Game-Theoretical principle of seeking a Nash Equilibrium. Since there are two Nash Equilibria, this decision-rule gives them no way to decide. But since our agents ultimately aim to bring about the best outcomes, shouldn't that lead them to prefer the payoff-dominant Nash Equilibrium – the Hi result – to the Lo result? But *this* outcome is not an option for either of them individually – it is not an outcome that either can bring about alone. And Act-Utilitarianism tells agents to choose only between options that are open to them individually. The only options available to each agent are to play 1 or 2, and the value of each of these is entirely dependent on what the other chooses. So long as the only thing that they each know about each other is that they are rational Act-Utilitarians, they cannot know what the other will choose: A is wondering whether he should play 1 dependent on whether B will play 1, but B will only determinately play 1 is she is sure that A will play 1, and if A is himself unsure that he should play 1, he must admit that B *cannot* be sure of this either – and vice versa.

You might think that Act-Utilitarians can solve this problem by applying the:

> *Principle of Indifference*: Where A has no idea what B is going to do, she can assign equal probabilities to each of B's options – as though B is going to choose at random.

In that case, A will assign a 50% chance to B choosing 1, and a 50% chance to B choosing 2. So now A can assign an expected utility score to each choice of hers – Option 1 has an expected utility of 10, and Option 2 has a score of 5. And so, as an Act-Utilitarian she should rationally choose Option 1. And by the same rationale, B will also choose 1.

This strategy has been rejected by many game theorists.[2] But, in any case, it yields the opposite conclusion in *The Altruist's Dilemma.* In that case, the Hi outcome requires *a great many* of the relevant agents choose to pursue political reform; if that will not happen, then it is better for all of them to devote themselves to charity instead. If each agent treated the others as having a 50% chance of pursuing either politics or charity, then each would expect that not enough others would pursue politics to make it worthwhile, and so would rationally conclude that the best thing would be to devote themselves to charity instead. In other words, if each agent, in the absence of any other basis for predicting how the others will act, applies the principle of indifference, then they will, in *The Altruist's Dilemma,* coordinate on the Lo outcome.

Thus, *even in the case where all the relevant actors are rational Act-Utilitarians*, DBWC's advice is sound Act-Utilitarian reasons really do *not* recommend that participants in *The Altruist's Dilemma*

[2] Colman and Gold (2020) and Bacharach (2006), argue that it is fallacious, because it involves a self-contradiction – it starts by assigning to each agent a 50% chance of playing each of A and B, and then concludes that they each have a 100% chance of playing A.

pursue political reform. If the Optimific Reasons are indeed reasons to pursue political reform, then the Optimific Reasons are *not* Act-Utilitarian Reasons – so the *Equivalence Thesis* is false.

9. Optimific Reasons and Cooperative Reasons

So we've now seen that ideally rational Act-Utilitarians cannot expect one another to pursue political reform, and so pursuing political reform is not the option with the greatest expected utility for any of them. So it is fortunate that, in real life, DBWC's audience are not likely to be idealised Act-Utilitarians. We've already seen that, if it could, DBWC should try to get its audience to pursue politics. But we've also seen, via the *Transmission* principle, that if DBWC has reason to get its followers to pursue politics, then they have reasons to do so. We now know that these reasons cannot be Act-Utilitarian. So what can they be?

Game theorists like Bacharach attempt to solve the puzzle of the Hi-Lo Game by appealing to the theory of "Team Reasoning". This is a theory of subjective rational decision-making, and so is not directly applicable here. But we can translate this idea into the language of objective moral reasons:

> *Team Reasons:* In coordination problems where multiple rational, moral agents can coordinate to bring about an optimific outcome, each agent has reasons to play her part in bringing about the best outcome that the collective of agents can jointly enact.

If agents follow their Team Reasons, then each of them will choose to pursue political reform, collectively bringing about the best outcome. So it is tempting to conclude that Optimific Reasons just *are* Team Reasons – at least in cases like *The Altruist's Dilemma*.

10. The Problem of Higher-Order Coordination

But, as Henry saw, this conclusion is also unsatisfactory. After all, it would be best for each agent to pursue politics *only if others were to do so.* Act-Utilitarian Reasons would not direct any agent to do so unless it was *already* the case that others were going to do so – it told them all to cooperate only *conditionally,* and a group of strictly conditional cooperators lacks a trigger to initiate cooperation. Team Reasons were introduced to break that impasse. They determinately tell agents to cooperate to bring about the best option that can be brought about in this situation. But, of course, the existence of *reasons* to pursue politics doesn't *force* people – even rational, moral people – to pursue politics! If enough of Henry's peers didn't realise – or didn't agree – that they had Team Reasons to pursue politics, then it would be sub-optimal if Henry pursued politics regardless.

In other words, we face a higher-order coordination problem. Just as the question of whether Henry should pursue political reforms depends on whether others are going to do the same, now we can see that the question of whether Team Reasons are Optimific Reasons depends on whether other people are *in fact* going to follow them. The problem can be stated as follows:

> *Higher-Order Coordination:* In a coordination situation, Team Reasons are Optimific Reasons if enough people are going to follow Team Reasons; if not enough people are going to follow Team Reasons, then Act-Utilitarian Reasons are Optimific Reasons.

It is tempting simply to fall back on the Act-Utilitarian approach of trying to *predict* what others are going to do in order to decide what I should do. And certainly, this sort of prediction does seem apt when a single agent be quite sure that others, due to immorality, stupidity or sheer pig-headedness, are not going to behave cooperatively, and perform the actions recommended by Team Reasons. If I am the only one interested in and capable of acting morally, it seems appropriate simply to predict what the others are going to do, and respond accordingly.

But in cases where several morally-motivated rational agents are on the stage, then if they all bring an Act-Utilitarian strategy to bear on the *Higher Order Coordination Problem*, they will simply reprise the problem of the Hi-Lo Game. If each agent will follow Team Reasons only if they predict that others are already going to, and if each agent knows that the others are thinking likewise, then there is no basis for any of them to make this prediction, and hence none of them will follow Team Reasons. And thus they will, in *The Altruist's Dilemma,* default to the sub-optimal equilibrium, and all focus on charity instead.

On the other hand, if they were guided by Team Reasons in *determining* which reasons to follow, then they would decide that the best thing they could do together was to *all* follow Team Reasons, and thus play their part by doing just that, and hence pursuing political reform. In other words, the *Higher Order Coordination Problem* seems just as intractable as the initial coordination problem of *The Altruist's Dilemma*. One way of thinking leads agents to Team Reasons and political reform, and the other leads agents to Act-Utilitarian Reasons and charity. Neither set of reasons seems to be determinately optimific until *after* the agents have acted.

But if that were true, then the principle that we should act according to optimific reasons could not guide agents *in* acting. And that is precisely when we need reasons! After all, as I have presented the matter, the only Utilitarian interest in providing a theory of reasons is in helping us to bring about better outcomes – a theory of reasons that could only be used to assess agents for their righteousness after the fact could be of no intrinsic interest.

11. The Dualism of Moral Reasons

So we need to revisit the construal of Optimific Reasons. My original formulation was:

> *Optimific Reasons:* Morality assigns to agents the reasons that it is best for them to follow; the optimific moral reasons are the reasons that will guide agents in such a way that leads to the best outcomes.

This formulation quantifies over agents in an ambiguous way. In the Altruist's Dilemma, it might mean *all* the relevant agents – which is to say, the entire audience of DBWC. In that case, the Optimific Reasons are Team Reasons. But it could also mean, *whichever agent is asking the question*. If Henry is wondering what he should do, and he predicts that not enough of his peers will follow Team Reasons, then the Optimific Reasons for *him* are to respond to the predictable behaviour of his peers by pursuing charitable giving – that's to say, the Optimific Reasons for him are Act-Utilitarian.

I think that *both* readings are legitimate. In other words, there are two quite different senses in which reasons might be optimific.

> *Outcome-Optimific Reasons*: In a situation where there is a best outcome that can be collectively brought about by moral, rational agents, the optimific reasons are the ones that direct these agents to bring it about.
>
> *Prediction-Optimific Reasons*: In a situation where there are objective bases to make predictions about the actions of agents, the optimific reasons for each agent are the ones that lead them to take the options with the greatest expected utility, given what can be predicted about the actions of all the other agents.

Both Act-Utilitarian *and* Team-Reasons are optimific *in one sense of optimific*. So the Equivalence Thesis is false – Act-Utilitarian Reasons are not identical with Optimific Reasons in *both* senses. But Team Reasons are not identical with Optimific Reasons in both senses either. In other words, I think there is a dualism of Utilitarian reasons – not Sidgwick's "Dualism of Practical Reason" between egoistic and moral reasons, but a dualism among *moral* reasons, between

Team Reasons that tell us to choose between options that are collectively open to groups of moral agents, and Act-Utilitarian reasons that tell each agent only to choose between options open to *her*.

12. What Explains the Dualism?

I now want to try to say something deeper to explain how there could be such a dualism of moral reasons. You might think that *Prediction-Optimific Reasons* are the only reasons that a Utilitarian should concern herself with. After all, our interest is in *making this world a better place*, not in performing acts that *would* be useful in an imaginary world of rational perfection.

But while this argument suggests that *Prediction-Optimific Reasons* must be *one* species of Optimific Reasons – and thus that Act-Utilitarian reasons are genuine expressions of *Meta-Deontic Utilitarian* commitments – the argument just given against *Outcome-Optimific Reasons* is too fast.

Henry, in his consideration of *The Altruist's Dilemma*, articulated a simple version of this argument:

> *Why Can't We Do What's Best?* If there is a best outcome that can be collectively brought about by moral, rational agents, the only thing that would stop them from achieving this best outcome is their own decision not to cooperate. And, so far as they are morally rational, the only thing that would make them decide this is if there were no reasons to cooperate. Since it would be best if they cooperated, there must be reasons to do so.

It seems hard to square with *Meta-Deontic Utilitarianism* that there might be case where the only obstacle to the best outcome obtaining is morality!

More deeply, Henry's thought shows the complex relationship of reasons to the causal structure of the world. The argument I gave against *Outcome-Optimific Reasons* assumes what we can call, following Bernard Williams (Smart & Williams, 1973), the:

> *Reactive Concept of Reasons*: What reasons a decision-maker has is fixed by all the other facts in the causal nexus. Questions about what can be predicted are *prior* to questions about what choices would be best in the determination of what an agent has most reason to do.

But this view treats each individual decision point as somehow special, detached from the nexus. It portrays the situation as though this one choice is open to rational guidance, whereas every other person's decision (and every future decision of one's own) is just a brute fact of nature. But if reasons can guide my actions, then they can also guide other people's actions, and so they partially determine, if defeasibly, what happens at other points in the causal nexus. So we should instead accept:

> *The Reciprocity of Reasons and Facts:* Optimising reasons attempt to guide agents to bring about the best outcomes, reacting to the facts in the causal nexus about what is going to happen; but they also guide other points in the causal nexus at the same time. Neither deliberation nor prediction can be strictly prior to the other.

Talking about the role of reasons in shaping the world in this way can sound mysterious – after all, reasons surely do not *cause* anything! But for the Utilitarian, this can be given a strictly naturalistic spin. After all, "reasons" in our sense are grounded in the fact that something is best, relative to a set of impartial desires. It is a standard part of belief-desire psychology that creatures often do things *because* they believe that these things best fulfil their desires. And these beliefs attempt to track objective facts. So it makes sense to say that creatures sometimes do things because they are best – and hence, that some facts in the world are determined by what reasons there are.

But while beliefs about what is best attempt to *react* to the causal nexus, they also *determine* parts of the causal nexus. In a world where there are many agents trying to do what is best, there is often no single determinate fact of the matter about whether their beliefs are true or not, because they are all simultaneously and reacting to the world and shaping it at the same time.

In other words, I think that my proposed Dualism of Moral Reasons follows from a deep fact about the nature of optimising thought in a world where there are many agents attempting to optimise. When we are all simultaneously reacting to and shaping the world, there is often no unique fact of the matter as to what option is best.

13. Overcoming the Dualism through Social and Political Change

This might seem to undermine the entire project of Utilitarian deontology, as I have portrayed it. We want reasons to guide us towards the best outcomes – but if the guidance of reasons is ambiguous, what use is it?

However, I think this would be unduly pessimistic. For Act-Utilitarian Reasons and Team Reasons can *align* in their prescriptions. Suppose that DBWC were able to make its audience unconditionally cooperative. This would give each agent Act-Utilitarian Reasons to cooperate as well. If Henry can *predict* that his peers will pursue politics, then, for him, pursuing politics is also the option with the greatest expected utility.

I argued that this could not help Act-Utilitarians. If DBWC had reason to make them more cooperative, then it must be true that they already had reasons to be cooperative, and Act-Utilitarianism denies this. But now that we have Team Reasons on the table, this problem falls away. If DBWC were to try to make its audience into default cooperators, then it would be trying to get them to do what they had *Team Reasons* to do. And *once* it had achieved this, it would also be true that they had Act-Utilitarian reasons to cooperate as well. If William or any of his peers were to deliberate by reference to Team Reasons, they would cooperate; but if any of them were to revert to an Act-Utilitarian framing, they would still cooperate. If only we can *become* default cooperators, then we have both Team- and Act-Utiliarian Reasons to continue cooperating. And Team Reasons recommend that we become cooperators. After all, becoming cooperators is the best thing that we, together, can do.

But being a *default cooperator* is different from deliberating by reference to Team Reasons on a case-by-case basis, even if default cooperators are acting *in accordance* with Team Reasons. But we can now help ourselves to the distinction between Objective and Subjective reasons. It's not required that agents always *think* in terms of Team Reasons when they act. Indeed, since those who deliberate might vacillate between thinking in terms of Team Reasons and Act-Utilitarian reasons, it might often be better if they do not deliberate in optimising terms at all. Rather, they should inculcate virtues, motivations, principles and rules into themselves that make them tend to do what they have Team Reasons to do.

These principles would not be rule of thumb attempting to approximate *Act-Utilitarian* reasons, as on the traditional view, but rather helping them to follow *Team Reasons* – and, moreover, helping them to do this resolutely and consistently, so that others could rely on them to play their parts. And it would *not* always be Rule-Worship to follow one's principles in cases which predictably lead to suboptimal outcomes if we thereby do what we have Team Reasons to do – since Team Reasons are one kind of Optimific Reasons, and we already know that following them can sometimes lead to suboptimal outcomes. Thus, the role of "Secondary Principles" is far more robust for the Dualistic Utilitarian than for the Act-Utilitarian.

Of course, developing and proselytising such principles and standards of virtue just is the work of social reform. Moreover, given the more robust role of these principles on my proposed Dualistic framework, there should be no need for the ultimate Utilitarian principles underlying these principles to be "esoteric". After all, the thought that the ultimate purpose of social norms and rules is to help us *cooperate* is a commonplace in many moral and political traditions, who do not feel the need to keep this thought a secret.

And there are more distinctively *political* aspects to this project. Perhaps the chief virtue leading us to act in the ways recommended by Team Reasons is that of *solidarity*. One of the lessons learned by the Trade Union movement and other radical political movements is that political collectives fighting towards a shared goal may do better when individual members refrain from individualistic strategic reasoning – that's to say, from being motivated by distinctively Act-Utilitarian Reasons. Utilitarian Dualism accommodates this – solidarity just *is* a virtue that makes following Team Reasons a default.

And finally, the project of aligning Act-Utilitarian- and Team- Reasons can make sense of programmes of governmental and legal reform. Bentham thought humans were psychological egoists, so that the role of law and public policy was to align egoistic motives with the greater good. He was undoubtedly too pessimistic in this view. But even if the citizenry were all altruists, there could still be a role for law and policy to align the individualistic altruism of Act-Utilitarian Reasons with the collectivist altruism of Team Reasons.

Problems like *The Altruist's Dilemma* – where we must choose between doing our individual best to try to mitigate the worst effects of a bad system, thereby failing to reform it and even allowing it to limp on with the aid of individual charity, or to risk trying reform it – are a symptom of political dysfunction. The early Utilitarians emphasised education, reasoned debate, free speech, equality, democracy, and latterly, cooperative socialism, as both a goal of political reform *and a means to effect it.* In a more enlightened democracy, we can hope, the best option for agents (outside the sphere of personal benevolence) is simply to do their part in sustaining and improving the functioning of an efficiently benevolent state and civil society.

Finally, and more darkly, we can see laws as a means to align Act-Utilitarian- and Team-Reasons. It is a commonplace that the law will punish those who violate its strictures, even if the individual acts in question are justified in Act-Utilitarian terms (except in some cases of great exigency). The alignment of Team- and Act-Utilitarian- Reasons requires that following Team Reasons be the default for all agents in the community. Thus, so long as the laws are reformed so that they are generally optimific, as Bentham hoped, then it makes sense to discourage Act-Utilitarian deviations from the path selected by Team Reasons – even if, in one important sense, deviators were following reasons that were, in their situation, optimific.

14. Conclusion

I have tried to do two things in this paper. The first was to challenge the supposed primacy of Act-Utilitarianism in the family of Utilitarian moral philosophies. Act-Utilitarianism is not the best version of Utilitarianism, even given purely Utilitarian premises. Rather, there is a Dualism of Utilitarian Moral Reasons – both Team Reasons and Act-Utilitarian Reasons deserve equal footing in Utilitarian ethics. There is more than one way to do our best.

This in turn vindicates two features of the early Utilitarian tradition. The first is the failure to state the "Principle of Utility" as an unambiguous theory of reasons, and the apparently excessive stress the early Utilitarians placed on "secondary principles". If my arguments are correct, then Utilitarian reasons *are* ambiguous, and principles deserve a greater status than contemporary Act-Utilitarians give them. The second is the early Utilitarian emphasis on social and political reform, rather than on personal morality. If my view makes sense, then such reform may be the best way to *make* the recommendations of morality unambiguous.

As to poor Henry – what should he do? In one sense, the answer is genuinely unclear, and my theory aims to explain why this is so. But perhaps one thought might tip the scale in favour of political reform. For, with luck, perhaps political reforms can help to create a world in which Henry's successors in altruism do not face the kind of dilemma that bedevils him.

Bibliography

Bacharach, Michael (2006) *Beyond Individual Choice: Teams and Frames in Game Theory* (Princeton, NJ: Princeton University Press)

Berkey, Brian (2018) The Institutional Critique of Effective Altruism. Utilitas 30 (2):143–171.

Colman, Andrew & Gold, Natalie (2020) "Team Reasoning and the Rational Choice of Payoff-Dominant Outcomes in Games," *Topoi 39 (2):305-316.*

Gibbard, Alan (1965) Rule-Utilitarianism: Merely an Illusory Alternative? *Australasian Journal of Philosophy 43.*

Gold, Natalie & Sugden, Robert (2007). Collective Intentions and Team Agency. *Journal of Philosophy 104 (3):109-137.*

Hooker, Brad (2000). *Ideal code, real world: a rule-consequentialist theory of morality*. New York: Oxford University Press.

Lazari-Radek, Katarzyna de & Singer, Peter (2014). *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. New York: Oxford University Press.

Lyons, David (1965) *Forms and Limits of Utilitarianism*, Oxford: Oxford University Press.

Parfit, Derek (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Railton, Peter (1984) "Alienation, Consequentialism, and the Demands of Morality" *Philosophy and Public Affairs Vol. 13, No. 2. 134–171.*

Regan, Donald (1980). *Utilitarianism and Co-operation*. Oxford: Clarendon Press.

Schelling, T. (1960). *The Strategy of Conflict*. Cambridge MA: Harvard University Press

Smart, J. J. C. & Williams, Bernard (1973). Utilitarianism: For and Against. Cambridge: Cambridge University Press. Edited by Bernard Williams.

Srinivasan, Amia (2015) Stop the Robot Apocalypse. London Review of Books 37: 3–6

Sugden, Robert (2015). Team Reasoning and Intentional Cooperation for Mutual Benefit. *Journal of Social Ontology 1 (1), 143 – 166*

Woodard, Christopher (2017). Three conceptions of group-based reasons. *Journal of Social Ontology 3 (1):102-127.*