Utilitarianism and Alienation:
How Damaging is the Alienation Objection for Utilitarianism?

1. Introduction

Imagine a lizard, going about its business, when it is attacked by some predator. The predator manages to grab the lizard by its tail, but the lizard has an ace up its sleeve. The lizard drops its tail, leaving the predator a morsel as it scampers away, relatively unharmed. One might call this a sub-optimal outcome for the lizard, and I'm sure the lizard would agree. But crucially, the lizard survives.

The alienation objection to utilitarianism is seen by many, even those who advance it, to be similar to a predator's swipe against a lizard's tail. Certainly, the charge of alienation might give us reason to abandon utilitarianism as a method of deliberation, or so the argument goes, but we should not confuse utilitarian deliberation for the truth of utilitarianism itself. The sophisticated utilitarian should drop utilitarian deliberation the same way a lizard drops its tail, and in a similar manner utilitarianism will emerge largely unscathed.

I will argue that utilitarianism cannot so easily drop its method of deliberation as a lizard does with its tail, specifically when concerned with the alienation objection. Instead, if the alienation objection succeeds, utilitarianism can be proven false. To demonstrate this, I will first argue that the alienation objection cuts deeper than other similar arguments for utilitarian self-effacement, and then argue that utilitarians should accept a robust notion of ought implies can. Then, I will show that this notion of ought implies can, coupled with the deep self-effacement brought about by the alienation objection, generates a contradiction in scenarios where utilitarianism recommends that some individuals not hold utilitarian values. I will show that, if successful, the alienation objection poses a much greater threat to utilitarianism than previously thought.

2. Outline of the Alienation Objection

I do not wish to comment on whether the alienation objection succeeds - my focus is on how damaging it would be were it to succeed. However, it is useful to lay out what I see to be the strongest form of the alienation objection. This argument is taken largely from Peter Railton's *Alienation, Consequentialism, and the Demands of Morality*, although elements are also lifted from Bernard Williams' *Against Utilitarianism*.

The first premise of my reconstructed alienation objection is as follows:

P1. Utilitarian values and partial values come into conflict.

Utilitarianism, by its definition, cares only about whether an action promotes happiness, not the roles of nor the relations between the persons involved in the action. A partial value on the other hand gives specific weight to certain actions and outcomes because of the unique situation and roles of the holder of the value. For example, if a mother was choosing whom to save in a life or death situation, and one of the potential options was the mother's child, utilitarianism states that the fact that one potential victim is the child of the mother has no inherent relevance to the moral decision making. After all, the only measure of an action's moral value, according to utilitarianism, is how much happiness the action produces. Therefore, nothing other than the strict, impartial sum of total happiness can recommend an action on the grounds of utilitarian values. But the partial value of being a mother would urge the mother to save her child, precisely because it is *her child*.

Utilitarians acknowledge that the mother's grief at losing her child may add weight to one side of the equation, and therefore say that the partial value of motherhood holds some

weight, but it holds no inherent weight. Utilitarian values do not care more about the mother's sadness *because she is a mother*, her sadness is equal to that of any other person in any situation. The mother has no commitments to her children over and above the commitments to her children's happiness, and her own, that every other person has. Thus, utilitarian values are still not partial: even if they can factor in the sadness or happiness caused by partial values, they cannot account for the force of the partial values themselves.

Partial values concern themselves with specific goods, not overall happiness. As such, partial values recommend courses of action that may conflict with what would produce the greatest happiness. The role of being a parent, for example, may guide one to care more about one's own children than they do other people's, even if this is to the detriment of the total overall happiness. It might be better for everyone if parents cared for others' children just as much as their own, but the partial values of parenthood require us to focus on our own children.

This conflict between partial and utilitarian values need not be relegated to relationships: partial and utilitarian values can conflict even when the partial value is directed inwards. Suppose one views oneself as a writer, indeed to the extent that one values writing as a part of one's self-identity: of what makes one's life worth living. Such a value could be described as a personal project by Williams. The writer would be compelled by their valuing themselves as a writer to devote a certain amount of time to developing their skill and producing books. But unless these books are incredibly good, and produce a great deal of happiness in everyone who reads them, the pursuit of writing will not produce as much happiness as say, volunteering at a charity, or even working extra hours and donating the money to charity. The partial value therefore conflicts with the utilitarian value.

But partial values and utilitarian values can still come into conflict even when they agree on outcome. Railton describes a situation where a husband is particularly loyal and good to his wife. But when he is questioned on why he cares for his wife's needs so much, he gives an abstract, utilitarian answer. He argues that he is in the best place to look after his wife's needs as he has a greater knowledge of them than other people, and that if everyone looked after their spouses in the way that he did, the total happiness of the world would increase. In this instance, the partial value of being a husband and the utilitarian values are recommending the same course of action: that of being a caring husband to one's wife, and yet the *reasons* for their actions are completely different. The husband in this scenario seems not to care about his wife because she is his wife, he cares about her because she is a subject of happiness and sadness, like everyone else. Thus, partial and utilitarian values can come into conflict not just in terms of what we actions they recommend we take, but also in the reasons for action, and the descriptions we act under.

The second premise of the alienation objection is:

P2. If a utilitarian and a partial value come into conflict, the moral value should overrule the non-moral.

This element of moral trumping - that moral concerns should trump any non-moral concerns - is generally understood to be part of what makes a value a *moral* value. We can understand that the values of motherhood, of passion, of friendship, both can and indeed should be given up when a value of greater strength comes along. Utilitarianism purports to be a system of ethics, and therefore utilitarian values are moral values. Partial values make exceptions of ourselves, and do not attempt to guide everyone equally, and we normally understand ethical values to be general. Therefore, utilitarian values, as moral values overrule partial values, as non-moral values.

For example, if I were to enjoy going for bike rides at 3 in the afternoon, but had promised that I would help a friend move a sofa at the same time, we would understand that my non-moral value of liking my 3pm cycle ride is trumped by the moral consideration that I should

keep my promise. But moral values are not so easily overturned. To return to the example of the writer I outlined above, the writer may decide to spend an afternoon composing sonnets instead of working at the local charity shop, but they cannot consider this decision to be the correct one if they hold utilitarian values and consider those values to be moral ones. Moral values trump non-moral values in the sense that one should always do the moral thing, and side with one's moral values.

Some might say that the normative not predictive sense of premise 2 weakens the argument. Our utilitarian values should overrule the others, they might say, but because they practically speaking do not, the overruling is not something we need to worry about, nor is something that any conclusion of substance can be built from. However, if people hold utilitarian values and understand them to be moral values, then they will let their utilitarian values overrule others, as that is what it means to hold the utilitarian value. The only ways that the utilitarian value holder would not let this overruling happen is if they did not hold their utilitarian values to be moral, and instead view them as one personal commitment amongst many, or if they acted akratically. But given that utilitarianism purports itself to be a moral theory, and akrasia by definition conflicts with one's better judgment, we may assume that the normative overruling of non-moral values by moral ones will result in a practical overruling, insofar as we might discover a contradiction within utilitarianism. If utilitarianism is spared a contradiction only because it is impossible to treat it's own values as truly moral, then utilitarianism is in a bad shape anyway.

It is also important to note that this moral overruling occurs in motivation as well as in action. Where morality is concerned, we must act for the right reasons, as well as performing the right actions to begin with. Consider two people who both save someone from drowning, but one does it out of a sense of moral righteousness, and the other does it out of hope for a reward. We would say that the former acted more morally than the other, because even though their values of wealth and moral values coincided, they acted because of the moral values. Thus, moral values trump non-moral values in terms of motivation, as well as in action.

The third premise is:

P3. If a value is overruled, then the holder of that value is alienated from it.

Railton defines alienation from a value as that value only featuring in one's feelings, not featuring in one's decision making or cognitive capacities. To be alienated from a value is thus to hold it but for it to make no impact on our decisions. A value being overruled by another does exactly that: we still hold the overruled value, it has some sway over us on an emotional level, but the value doesn't result in any action. We don't even deliberate on the value because we know another value simply trumps it, the strength of the respective values regardless. Consider again the husband in Railton's scenario: he is alienated from loving his wife because the love he feels, if he feels any at all, does not play a part in his decision making. The utilitarian values of maximising happiness justify his actions, not the love of his wife. The love he feels for his wife won't ever impact his actions as long as the utilitarian values hold. Thus, they will never feature in his deliberation: they are not a part of his practical reason or cognitive capacities. Therefore, he is alienated from them.

Now, we reach the conclusion:

C1. Utilitarian values alienate their holders from their non-utilitarian values.

Utilitarian values will come into conflict with our partial values. But because utilitarian values are meant to be moral values, these values will overrule the non-moral partial values. But because they are overruled, the partial values no longer take an active role in being action guiding, or even featuring in deliberation. We may feel their pull, but we should never act on them, or even be motivated by them. Thus, we are alienated from them.

The full argument is thus:
P1. Utilitarian values and partial values come into conflict.
P2. If a utilitarian and a partial value come into conflict, the utilitarian value overrules the partial.
P3. If a value is overruled, then the holder of that value is alienated from it.
C1. Utilitarian values alienate their holders from partial values.
So where does this leave utilitarianism? Utilitarian values may alienate us from non-utilitarian values, but what does this mean for the theory as a whole? Well, Railton argues that just as there is a paradox of hedonism, so does alienation reveal a paradox in utilitarianism. Consider the conclusion of the above argument as the starting premise of a new argument. Utilitarian values alienate their holders from non-utilitarian values. So what? Well, alienation from one's values is a state of unhappiness, or at the very least is counter-productive to happiness. One's partial values are what determines what we desire in the first place, and therefore determine what happiness amounts to for us. Being alienated from the source of our own happiness is therefore counterproductive to the pursuit of the greatest happiness for the greatest number. Certainly, it might not outweigh other concerns: it is easy to imagine that the lives of others are of greater import in the utilitarian calculus than someone's alienation, but for some set of people, the best way for them to maximise their happiness is to not hold utilitarian values.
Exactly how large this set of people will be depends on how bad the harm of alienation is, which I will leave to be an open question. Certainly if you follow Williams' view that our ground projects are the source of our happiness then alienation is a very great harm indeed, and therefore in most circumstances happiness will be maximised by not holding utilitarian values. One could reasonably claim that the harm of alienation is less than cutting oneself off from the possibility of happiness, but the fact that alienation distances us from the source of much of our happiness: our partial values, means it should be given some serious weight in terms of the utilitarian calculus. Additionally, Railton's scenario of the husband shows that alienation doesn't just harm our own happiness, but the happiness of others. The husband might be a perfectly caring spouse, but if his wife finds out about his true, wholly utilitarian motivations, she may feel hurt. Therefore, for some set of people, it will increase overall happiness more if they are not alienated than if they adopted utilitarian values.
But utilitarian values themselves claim that we must do what will maximise happiness, and as we have just established above, for some portion of people that is to give up utilitarian values. Utilitarianism therefore recommends that we not be utilitarian, or at least recommends that some of us not be utilitarian, in order to maximise happiness for the greatest number. To put the above in premise-conclusion form:
P1. Utilitarian values alienate their holders from their non-utilitarian values.
P2. At least for some people, in some situations, alienation would cause great enough personal unhappiness (and they would be inept enough at creating happiness for others were they alienated) that to maximize happiness, they must not hold utilitarian values.
P3. Utilitarianism recommends that we maximise happiness.
C1. Utilitarianism recommends that, for some people, in some situations, they should not hold utilitarian values.
Again, the purpose of this essay is not to comment on the soundness of the above argument. Whilst the argument is deductively valid, and its premises do not seem wholly controversial, I do not deny that they are open to challenges. What I am concerned with here is what this objection would mean for utilitarianism, *were it to succeed.* We have here an argument for rejecting utilitarian values on utilitarian grounds. We can say therefore that utilitarianism is self-effacing: it recommends not following its own precepts. But is this self-effacement a

problem for utilitarianism, or can it even be seen, as Railton and others argue, as a positive?

3. Benign Self-Effacement

In order to explain how some claim utilitarianism as a theory can survive this self-effacement, we must first explain a distinction Railton makes between subjective and objective utilitarianism. Subjective utilitarianism is in essence utilitarianism as a method of deliberation. Those who adopt subjective utilitarianism will perform the action that they believe will cause the greatest happiness for the greatest number. For example, a subjective utilitarian would approach a situation such as deciding whether to shoplift a Mars bar by weighing up the harm that stealing a Mars bar and the accompanying loss of funds for the company, with their personal happiness of not having to pay, and a nice Mars bar.

Objective utilitarianism on the other hand is the belief that the criterion for a good action is whether it promotes the greatest possible happiness for the greatest number, but it doesn't necessarily follow from such that the greatest happiness can be accomplished by conscious utilitarian deliberation. Suppose it was discovered that, as a point of psychology, humans are terrible at calculating the expected happiness of actions, and it is much more reliable for them to adopt rules of thumb. The objective consequentialist would know that more happiness will be achieved by following such rules of thumb, and so would abandon subjective utilitarianism for a rules-based method of deliberation.

The distinction between subjective and objective utilitarianism, and the fact that the two can come apart, is what allows Railton to claim that the alienation objection does no harm to utilitarianism itself. If utilitarian values are counter-productive to utilitarian aims, then the sophisticated utilitarian is well within her means to drop subjective utilitarianism entirely as a method of deliberation. The sophisticated utilitarian can adopt any method of deliberation they like as long as that method is likely to produce the greatest happiness for the greatest number of people. The self-effacement utilitarianism faces under the alienation objection is therefore benign. Objective utilitarianism is unscathed even if it denies subjective utilitarianism.

This is the sense in which utilitarianism can be said to be like a lizard. The lizard's tail is like subjective utilitarianism: it seems to follow necessarily from the rest at first glance, but the rest can do perfectly fine without it, and in fact better in some cases. The alienation objection therefore is only a problem for subjective utilitarianism: there is nothing in it that casts doubt on utilitarianism as a criterion for right action.

Or at least, nothing directly in the argument that casts doubt on objective utilitarianism. But if the kind of self-effacement caused by the alienation objection throws up problems for other principles in utilitarianism, then the alienation objection will have revealed some inconsistencies in objective utilitarianism itself. In the next section I will discuss exactly what kind of self-effacement is caused by the alienation objection, and why it is much more problematic than other kinds of self-effacement.

4. Shallow and Deep Self-Effacement

Railton lays out, as examples for his theory of self-effacing utilitarianism, a few other ways in which utilitarianism might be self-effacing, and yet objective utilitarianism could still be correct. It is in these examples however that we find the first problem for Railton, as he seems to consider these examples as all roughly of the same kind of self-effacement. However, there seem to be two distinct types of self-effacement emerging in his examples, and I shall attempt to distinguish them.

The first kind of self-effacement found in Railton are situations where deliberating like a utilitarian is merely impractical, such as when considering whether to have chicken or

beef for lunch. Say someone prefers beef markedly, but also acknowledges that chicken is better for the environment than beef is. Then again, they might reason, the increased demand for chicken might be worse than the demand for beef, and so on and so on. The utilitarian impact of this person's choice is minor, but the calculus they are performing is quite complicated, and it is arguable that there are better uses of their time than utilitarian deliberation. Picking a rule of thumb for themselves will increase overall happiness more than doing the calculus each time, and so we have a self-effacing utilitarianism. This self-effacement we may call shallow self-effacement, as there is nothing in the interests of time that stops someone from holding utilitarian values. The indecisive lunch picker can absolutely make it their aim to increase the greatest happiness for the greatest number of people, and hold that value as their chief goal, but yet use rules of thumb as a sophisticated utilitarian. Shallow self-effacement therefore would never require someone not to hold utilitarian values.

The second scenario in Railton that I would like to discuss is the Kantian Demon. Imagine that there is a demon secretly controlling the world, and yet this demon is also committed to the cause of Kantian ethics. Not committed enough to actually be a Kantian, but committed enough to cause abject harm and misery to anyone that engages in consequentialist thinking. If someone holds utilitarian values, then the demon will make sure their life is a misery. The demon only leaves alone those who are Kantians. If there was such a demon, then a utilitarian who was committed to the maximisation of happiness for the greatest number would have a very good reason to ensure that most people were not utilitarians. More than this however, they would have a good reason to make sure that most people wouldn't even think like a utilitarian or hold utilitarian values, as doing so would deny them their happiness. If they themselves were more likely to maximise happiness by avoiding the Kantian demon's wrath, then they would have a utilitarian reason to rid themselves of all utilitarian thoughts. We can call this second kind of self-effacement deep self-effacement, as instead of merely denying that we should *deliberate* like a utilitarian, deep self-effacement denies that we should *hold the values* of one either. It denies the adoption on a deeper level: on the level of values, instead of practical considerations.

Which kind of self-effacement does the alienation objection give us then? Is the worry of alienation a merely practical consideration, such as with time constraints, or does the worry of alienation more resemble a Kantian demon? Railton seems to suggest that utilitarian values and partial values can co-exist in the same person: as he sketches in the character of Juan. Juan is a doting husband like in Railton's previous example, but when asked why he dotes on his wife, he replies that it is because he loves his wife. He acts out of a genuine partial value: the love of his wife, and yet seems to fit this partial value in the rest of his utilitarian worldview. Therefore, the alienation objection would only yield a shallow self-effacement.

I disagree with Railton's assessment, however. Utilitarian values are in a sense caustic to others, in that as long as utilitarian values and partial values are both present, the former will erode the latter. If utilitarian values are caustic, they cannot happily co-exist with partial values, and so only deep-self effacement can save us from alienation.

5. Utilitarian Values as Caustic

It isn't unreasonable to assume that every situation a person will find themselves in will have multiple possible actions, and these outcomes will also vary in their levels of happiness. Thus, every situation will be relevant to utilitarian values, as they concern themselves with a ubiquitous feature of human experience: happiness. In addition, as long as we hold that moral values overrule non-moral values on principle, in every situation that utilitarian values are

relevant, they will overrule other values. Thus, in every situation, utilitarian values will overrule non-utilitarian values. To put this in premise-conclusion form:
P1. Utilitarian values are relevant to every situation.
P2. Utilitarian values overrule non-utilitarian values in every situation they are relevant.
C1. Non-utilitarian values are overruled in every situation.
Therefore, utilitarian values can be said to be caustic to non-utilitarian values, in that wherever both are present, the non-utilitarian value is eroded by the utilitarian one. The agent holding both is not justified in ever being influenced by their non-moral value due to the universal applicability of utilitarian values. They will therefore inevitably face alienation from their non-moral values as long as utilitarian values are also present. There is simply no scenario in which our partial values have a chance to get a word in edge-ways, as long as utilitarian values are universally applicable and automatically overruling.

To see this causticity in action, consider someone who values being a poet, and who also holds utilitarian values. There are two possible options for such an individual: either the course of action recommended for them by utilitarianism coincides with their desire to be a poet, or it doesn't. In the latter instance the utilitarian value overrules the value of poetry, and the poet is alienated from that value, as they are frustrated in their personal goals by the demands of utilitarian morality. But even in the former instance the poet is alienated. Let us say that utilitarianism commands them to write poetry, as the agent has figured out that their poetry will produce a greater happiness for all than any of the other projects they could have done that day. Even in that instance, the poet is being alienated, because they must do poetry not for the sake of their partial values, but for the sake of their utilitarian values. Recall from section 2 that the overruling of partial values also occurs in motivation, not merely in action. There is no scenario where utilitarianism will allow the agent to act on their partial values, thus the utilitarian value will eat away at any other value the agent holds.

If then, there is some set of people who can best maximise happiness by avoiding alienation, then these people must give up their utilitarian *values*, not just their utilitarian *deliberation*. It is not enough, as Railton suggests, to merely deliberate in a non-utilitarian manner whilst still holding the maximisation of the good as one's highest value. In Railton's example, Juan even explicitly admits that he is no saint: he acknowledges that he is acting against what utilitarianism tells him to do. If we do treat utilitarian values as moral however, and thus as overruling, we cannot do what Juan does. We must alienate ourselves from our partial values if we truly are to treat utilitarian values as moral.
Utilitarian values are therefore caustic to other values, and so if one wants to truly reap the benefits of partial values, they must rid themselves of their utilitarian values. They must learn to act on partial feelings, not universal morality, if they are to meet the aims of that universal morality. The alienation objection is therefore committed to a very deep kind of self-effacement indeed, where we do not even hold the maximisation of happiness as a value. Still, at this stage we do not have a problem for objective utilitarianism. Deep self-effacement is certainly an odd situation, but not problematic per se. The problems do begin to emerge, however, when we consider another principle that utilitarianism relies on: that of ought implies can.

6. Ought Implies Can, Intentionally

Eric Wiland's paper *How Indirect Can Indirect Utilitarianism Be?* argues that forms of utilitarianism that are deeply self-effacing run into a problem regarding the principle of ought implies can: the principle that if one *ought* to perform some action, one must also be *able* to perform said action.
Firstly, it is worth establishing that utilitarianism is committed to some notion of ought

implies can. Wiland notes that the core principle of utilitarianism: that of maximising happiness for the greatest number, already presupposes a notion of ought implies can. Maximising happiness means creating the greatest amount of happiness *that you can*, not the greatest happiness full stop. Consider Parfit's scenario where two people are in danger, and you only have the time to save one. Utilitarianism traditionally would say that you have an obligation to save the life that will contribute best to overall happiness. It does not claim that you ought to rescue both of them, even though you don't have time to do so, because that would be impossible. The fact that utilitarianism makes us choose between the two, and claims that saving one of them is an optimally moral action, points towards utilitarianism being committed to ought implies can.

Wiland however argues that ought implies can (hereafter OIC) implies the existence of another, related principle: that of ought implies can, intentionally (hereafter OICI). OICI demands that if we are obligated to perform some action, it must be possible for us to perform that action intentionally: as a result of our will and not in conflict with it. If I were obligated to save a child drowning in a pool, for example, according to OICI, I must be able to will to save that child. I must be able to save the child of my own accord, and as a result of my values. Note that the principle of OICI does not imply that I must be able to save the child given my current value set. Such a principle would be far too permissive, and essentially hand a blank chequebook to the selfish to do as they will. Nor is 'intentionally' here being used in the sense that one must consciously have one's intentions "running through one's mind" as Wiland puts it. Reducing intentional actions to actions dictated by occurrent thoughts would be much too restrictive. Rather, OICI demands that it must be possible to perform an action as a result of a value set, or to put it simply: if we are obligated to Φ, it must be possible for us to Φ whilst intending to Φ.

To motivate OICI, Wiland asks us to consider some action. Let us call it Φ. Let us also say that we are told we have an obligation to Φ. However, we are also told that by intending to Φ, we will ruin any chance we have at Φ-ing. Is it reasonable to say we have an obligation to Φ if we can only Φ unintentionally? We might be tempted to say yes, because we can attempt actions indirectly. Consider the command "don't think of pink elephants!". Whilst not thinking of pink elephants might be difficult for a split second, we can intentionally think of other things, and focus on them. Therefore, we can Φ, even though we can't Φ intentionally. But we are intentionally performing actions, in this scenario, that will ensure the success of Φ. We are still Φ-ing intentionally, just in a roundabout manner. If intending to Φ inherently scuppers our chances of Φ-ing successfully, these indirect strategies will not work.

Indeed, there is no way that we can possibly Φ whilst intending to Φ, so in what sense can we do it? One might say that we can Φ in the sense that it is entirely possible that we might Φ accidentally, while intending to do something else, but there is no scenario in which we can ensure that we Φ. If one is only able to Φ entirely by accident, we cannot say that, practically speaking, one can Φ at all. Unless one can ensure that one Φ 's successfully, it is practically impossible for one to Φ . Thus, if it is impossible to Φ, then we cannot be obligated to Φ, by the broader principle of OIC. Thus, OIC implies OICI.

Put in premise-conclusion form, Wiland's argument is:

P1. It is not practically possible to Φ, without intending to Φ.

P2. We can only be obligated to do what is practically possible.

C1. We cannot be obligated to Φ without being able to Φ and intend to Φ.

The issue with Wiland's argument is that it presupposes a practical reading of OIC in P2. There is an ambiguity at the heart of OIC regarding what notion of 'can' is being used: practical possibility or logical possibility. The logical possibility interpretation claims that we can only be obligated to do that which is logically possible, for example we cannot be

obligated to draw a square circle, or two add two and two together to make five. However, there is nothing in the logical interpretation of OIC that prevents us from being obligated to swim in order to rescue someone, even if we aren't able to swim. It is logically possible for us to swim, so we can be obligated to do so. A more practical reading of OIC however claims that we can be obligated to perform actions that are within our practical means. Under the latter interpretation, we could not be obligated to swim to rescue a child as it is practically impossible at that moment to do so. It might be logically possible, but is outside our ability to ensure that we swim across, and so we cannot be obligated to do so.

Wiland's argument needs the practical possibility reading of OIC, as without it our inability to Φ non-accidentally isn't a huge problem for OIC. It is still logically possible for us to Φ, in the circumstances where we Φ accidentally there is no logical contradiction, and so Wiland's argument isn't persuasive. In order for Wiland's argument to work, we need a reason to prefer a practical reading of OIC over a logical one.

Such a convincing argument for a practical reading of OIC can be found in Bart Streumer's *Reasons and Impossibility*: his argument from deliberation. Streumer argues that a practical reading of OIC can be justified by an appeal to a similar principle regarding reasons. We can only be obligated, Streumer argues, to perform actions that we have reasons to perform, however 'reasons' is cached out. The idea being that the strength of the reasons for the action is what makes it obligatory. Thus, if there is a limit on reasons, such that we only have reasons to perform actions that we can practically accomplish, then we can only be obligated to perform actions that we can practically accomplish. Streumer calls the principle that we only have reasons to perform actions that we can accomplish R.

Suppose then that R is false. Without a restriction on what reasons we have, we would have reasons to do all sorts of things. At any given moment, I have a choice of actions. I can continue tapping out this essay, I can get up and make myself another coffee, I can even start doing jumping jacks. I have reasons to do all of these, varying in strength. But if R is false, then I also have reasons to go back in time and kill Hitler, reasons to end all suffering in the world with a click of my fingers, reasons to in the next five minutes fix my house's plumbing. These potential actions are not outside the bounds of logical possibility, and so even if they are outside of practical possibility, I still have reason to do them. One might find the existence of such reasons argument enough to accept R, but the situation gets worse when one factors in deliberation. One has to deliberate on these reasons, and given the vast amounts of good that the impractical options will result in, we must give deliberative weight to them. Continuing to write this essay might be a good thing to do, but preventing all suffering in the world with a click of my finger is orders of magnitude better, and so should be the preferred action. The fact that we could never so deliberate, and in fact deliberating as such would be impossible, gives us reason to believe that R is true.

One might be tempted here to say that we do have such impossible reasons, it's just that we do not deliberate on them, and so impossible reasons do not pose the problem that the argument from deliberation accuses them of. But reasons are meant to have deliberative weight, otherwise they are not reasons. If these impossible reasons hold no deliberative weight, we might ask why we call them reasons, or even why we posit their existence in the first place. Thus, we should accept that R is true, and if R is true, then a practical notion of OIC is entailed.

To finally return to OICI, a practical reading of OIC necessitates OICI. One cannot be obligated to perform actions that are not practically possible for one to do, and being unable to ensure the success of an action without relying on luck is surely a breach of practical possibility. The unlucky person in Wiland's scenario cannot practically perform the action that is demanded of them because as soon as they intend to perform the action, they fail.

Thus, OIC entails, as long as we read OIC practically, OICI. Because utilitarianism has a commitment to OIC, it therefore has a commitment to OICI.

7. OICI and Deep Self-Effacement

Now we have both a deeply self-effacing form of utilitarianism, from sections 4 and 5, and a utilitarian commitment to OICI, from section 6, I can demonstrate why the alienation objection poses problems for objective utilitarianism, as well as subjective. Consider one of these people that ought not to hold utilitarian values. They ought not to hold utilitarian values because of objective utilitarianism: the morally correct thing for them to do is ensure the greatest happiness for the greatest number of people. If they are alienated from their values, then their ability to feel happiness will dip, and so as long as holding partial values and ridding themselves of utilitarian ones doesn't cause more harm than the harm of alienation, they ought to adopt a non-utilitarian, partial value set.

However, once these people have adopted a partial value set, the truth of objective consequentialism doesn't just go away. They are still obligated to maximise happiness, as according to objective utilitarianism, the criterion for a good action is still the action that promotes the greatest happiness for the greatest number. Therefore, the person who is obligated not to hold utilitarian values is at the same time obligated to promote the greatest happiness for the greatest number.

This is where OICI comes into play. The individual is obligated to maximise happiness. By OICI, the individual must then be able to maximise happiness *whilst intending* to maximise happiness. But in order to maximise happiness intentionally, they must both maximise happiness, and hold utilitarian values, as the utilitarian values are what makes the maximisation of happiness intentional. As stated above however, the alienation objection states that for this individual, holding utilitarian values will prevent them from maximising happiness, as the best way to ensure the greatest happiness for them is to ensure their own happiness, which would be ruined by their alienation from their values.

The individual has two and only two options: to hold or not to hold utilitarian values, both of which contradict utilitarianism. If the individual holds utilitarian values, then they will fail to maximise the good, as they will be alienated from their own values and, according to the scenario, this will cause more harm to overall happiness than good. They may have intended to maximise the good, but they failed to do so. If the individual doesn't hold utilitarian values and continues to use personal values in their reasoning, then they may end up maximising happiness unintentionally, but they cannot intentionally promote happiness as in order for that to be their intention, they must have utilitarian values. Crucially, one must both intend to maximise the good, and do so, and the prerequisite for intending to maximise the good (utilitarian values) is exactly what ruins one's chances of maximising the good. One cannot do both. The alienation objection therefore shows that in order to maximise the good, some people in the right circumstances must not have utilitarian values. OICI shows us that we must have utilitarian values, otherwise the dictates of utilitarianism violate OICI.

In premise-conclusion form, the argument is as follows:

P1. There is some person, X, for whom holding utilitarian values prevents them from maximising the good. (Deep self-effacement, sections 4 and 5)
P2. If one ought to Φ, then one can Φ intentionally. (OICI, section 6)
P3. Every person ought to maximise the good. (Definition of objective utilitarianism.)
C1. X can maximise the good intentionally. (P1, P2, P3)
P4. In order to maximise the good intentionally, then one must both maximise the good, and hold utilitarian values. (By definition of intentional action.)
C2. X cannot maximise the good intentionally. (P1, P4)

C1 and C2 are explicitly contradictory.
Objective utilitarianism, when combined with the deep self-effacement that the alienation objection brings, does generate contradictions, when its commitment to OICI is brought out. Even though in principle subjective and objective utilitarianism may come apart, if the abandonment of subjective utilitarianism makes the obligations of objective utilitarianism impossible, then objective utilitarianism is harmed by the abandonment of subjective utilitarianism. Thus, while on the face of it the alienation objection only targets utilitarian deliberation, it in fact poses a problem for utilitarianism as a whole.

8. Objections

The first objection my argument might face is that self-effacement is only a problem in a minority of cases. One might push that alienation is not that great a harm: perhaps lower pleasures can still be enjoyed without full access to one's partial values. Maybe it turns out that alienation doesn't completely stop us deriving happiness from our partial values, it only mildly frustrates them. If the harm that alienation causes is rendered minor enough, then the number of people who would be obligated by utilitarianism to abandon utilitarian values would be very small. Perhaps in most cases the utilitarian benefit of being able to evaluate the best course of action outweighs the harm of being alienated from one's partial values.

This objection has two responses. Firstly, we may question how useful utilitarian deliberation is for most people who are not regularly placed in ethical dilemmas. Most people live their lives in quite clear ethical situations where partial values that promote good rules of thumb would line up quite well with what utilitarian values would tell them to do. The only places where there would be significant differences between the partial value person and the utilitarian value person would be in morally complex situations that require a non-obvious utilitarian calculus, and those situations are generally rare. So even if the harm of alienation is minimal, the benefit of utilitarian deliberation is also minimal, so the set of people who would be better of ensuring their own happiness by avoiding alienation to promote the greatest happiness is not vanishingly small.

But even if it was a vanishingly small set, the contradiction I have laid out is a problem for utilitarianism, however often it comes up. Utilitarianism as an ethical system is a universally quantified theory: that for all people in all situations it is good to maximise happiness for the greatest number, and if there is a contradiction even in one case, it shows utilitarianism to be false. The alienation objection therefore shows that the value of maximising happiness, so long as happiness is negatively impacted, however slightly, by holding the value itself, is a contradiction for utilitarianism and its commitment to OIC.

The second possible objection is to ask why someone who indirectly manages to not think of pink elephants is any different to a former utilitarian who manages to rid themselves of utilitarian values. Both intentionally perform an action (either thinking of other things such as orange ducks or ridding themselves of utilitarian thoughts, respectively) and then succeed in performing another action that would, if directly intentionally performed, fail. If the not thinking of pink elephants is considered intentional, then surely the utilitarian's maximising of the good is intentional?

The difference between these two cases is that the former utilitarian can be said to have intentionally rid themselves of utilitarian values, but that doesn't mean that they intentionally maximise happiness. Once they are rid of utilitarian values, they may maximise happiness, or they might not, depending on their partial values. Either way, the utilitarian cannot ensure that they maximise the good by ridding themselves of utilitarian values in the way that the intentional forgetter ensures they do not think of pink elephants by actively thinking of orange ducks.

9. Conclusion

I do not believe the arguments above doom utilitarianism. The utilitarian can of course deny the premises of the alienation objection, or find some way of showing the argument to be invalid. But utilitarians can no longer accept the alienation objection and view it as non-harmful to objective utilitarianism, not without either pushing back against a OICI, or giving an account of how utilitarian values can co-exist peacefully with their partial value counterparts. To return to my analogy at the beginning, the alienation objection doesn't just take the tail of the lizard. If it succeeds, it eats the lizard whole.

Word Count: 7997

Bibliography

Harcourt, Edwin, '*Happenings Outside One's Moral Self' Reflections on Utilitarianism and Moral Emotion* (Philosophical Papers Vol. 42, No. 2, July 2013)

Lewis, David, *The Paradoxes of Time Travel* (Oxford University Press, 1976)

Mill, John Stuart, *Utilitarianism* (The Floating Press, 2009)

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

Railton, Peter, *Alienation, Consequentialism, and the Demands of Morality* (Philosophy and Public Affairs, 1984)

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

Streumer, Bart *Reason and Impossibility* (Philosophical Studies, Volume 136, 2007)

Wiland, Eric, *How Indirect Can Indirect Utilitarianism Be?* (Philosophy and Phenomenological Research, 2007)

Williams, Bernard, *Moral Luck* (Cambridge University Press, 1981)

Wolf, Susan, *Moral Saints* (The Journal of Philosophy, Volume 79, Issue 8, August 1982)

Zangwill, Nick, *Insane Consequentialism* (Utilitas, Volume 30, Issue 3, September 2018)