Contractualism and the conditional fallacy

JUSSI SUIKKANEN

Final author copy; To be published in *Oxford Studies in Normative Ethics*

I Subjunctivist Accounts and the Conditional Fallacy

Philosophers often give accounts of the actual properties of objects in terms of what would happen to them in hypothetical circumstances.[1] These 'counterfactualist' or 'subjunctivist' accounts take the following form:

(Basic Equation) X is F if and only if, and just because, if X were in the circumstances C, then X would be G.

Here, on the right-hand side of the biconditional, there is an embedded subjunctive conditional. Its antecedent places the object in certain circumstances C which may or may not obtain. If the object then has the property G in those circumstances as the consequent claims, then, as a result, the object has the property F in the actual circumstances.

Examples of such views include:

(Redness) X is red if and only if, and just because, if normal observers were to look at X in standard circumstances, then X would seem red to them.

(Goodness) X is good if and only if, and just because, if X were an option for fully rational and informed agents, they would desire X.

These appealing accounts are said to contain a structural flaw. The claim is that, as subjunctivist accounts, they commit the 'conditional fallacy'.[2] The crux of this objection is that, at least in some cases, placing X in the hypothetical circumstances C changes the nature of X in such a way that whether X is G in those circumstances will be irrelevant for whether X is F in the actual circumstances. In these cases, X is significantly different in the

[1] These accounts are typically discussed in the literature about dispositions and powers (see fn. 2). The question here is often about whether dispositions can be reductively analysed in terms of manifestation- or stimulus-conditions and the responses in these conditions.

[2] This discovery was originally made by Shope (1978), but see also Johnston (1992), Wright (1992, 117–120), Johnston (1993), Martin (1994), Lewis (1997), Bird (1998), Johnson (1999 and 2003), Fara (2005), and Bodevic et al (2006).

circumstances C when compared to what it is like in the actual circumstances, and therefore it would be fallacious to draw conclusions about X's Fness on the basis of what happens to it in C. In other words, placing X in C can alter X so as to interfere with the truth value of the claim that X is F (Wright 1992, 117–118). And, because of this, the instances of the Basic Equation are claimed to be unable to provide correct extensions for the analyzed properties.

This problem is easy to illustrate with the previous examples. Firstly, in the case of Redness, consider a shy Chameleon which always turns red from blushing when ordinary people are looking at it (Johnston 1992, 231). Intuitively, this Chameleon can be green when it sits on a leaf unobserved. However, even in that situation it is true that, if we placed the Chameleon in bright daylight for people to observe, it would look red. Thus, Redness claims that the Chameleon is red (instead of being green) even when it sits on the leaf unobserved. This absurd conclusion illustrates how placing the object in the circumstances C can change the object so as to distort the truth about its colour.

Goodness fails in the same way. Consider an emotional squash player who cannot shake hands with his opponent after a lost match without losing his temper and hitting her (Watson 1975, 210, Smith 1995, 111). Goodness entails that it would still be good for him to do so. This is because a fully rational version of him would want to shake hands with his opponent. Given his rationality (being cool, calm, and collected), he could easily refrain from hitting her. So, here too, considering the act in the counterfactual situation changes its nature so as to make the consequences irrelevant for the goodness of the act in the real world.

Philosophers have reacted to this structural problem in two ways. The first has been to seek a universal solution. This project has been pursued by trying to reformulate the Basic Equations so as to avoid the conditional fallacy. This strategy is evident in David Lewis' attempt to move from the Basic Equation to conditional analyses that are indexed more carefully to times and causal bases that are stipulated to remain the same in the counterfactual conditions (Lewis 1997), and in Crispin Wright's move from the Basic Equations to his 'provisional equations' of the form 'If circumstances C obtain, then X would be F if and only if X would be G' (Wright 1992, 119). I will not evaluate the costs and benefits of these moves here.

The second strategy has been to try to find local ways of avoiding the conditional fallacy in the contexts of particular properties. Michael Smith's move from Goodness to the 'advice-model' is a good example of this strategy (Smith 1995, 111). Smith draws a distinction between the 'evaluating' and the 'evaluated' perspectives. On his account then, an act is good in the evaluated actual circumstances if and only if, and just because, a fully

rational and informed version of an agent would, from her idealised circumstances, want the actual agent to do that act in the evaluated circumstances in the real world. So, in the previous example, the fully rational version would want the emotional squash player to avoid shaking hands with his opponent because he would know that in the actual circumstances the agent will not be able to avoid hitting her.

This move helps to prevent the distortion of the evaluated circumstances when we make the antecedent of the subjunctive conditional true on the right-hand side of the biconditional. Making the agent fully rational in the distinct evaluating circumstances cannot affect the actual evaluated circumstances. Because of this, the conditional fallacy can be avoided in this context. However, it is clear that this strategy does not work in all cases. It cannot, for instance, fix Redness. There are no fully rational versions of chameleons that could give advice for their actual versions.

In this chapter, I want to investigate the conditional fallacy in a specific local context. Different versions of contractualism have recently become very popular in normative ethics. It is true that some of these contractualist theories have an 'actualist' structure. According to these views, the moral status of acts always depends on some actual agreements that have been reached in the real world. Thus, for example, according to Gilbert Harman's contractualist view, whether some act is right or wrong must always be determined by an actual implicit agreement within a group that includes at least both the agent who is doing the act and the people who are evaluating it (Harman 1975).

One of the main problems with such theories is that, if the relevant group of people has failed to reach an agreement about the moral status of a given act, then that act is neither right nor wrong. Therefore, as people do not currently agree about whether having an abortion is morally right or wrong, no moral attributes could apply to abortion on Harman's view. For most people, this seems like an unacceptable consequence of the actualist forms of contractualism.

As a result, most contractualist views in normative ethics have recently had a counterfactualist/subjunctivist structure. Very roughly, these theories attempt to give an account of rightness and wrongness in terms of the moral principles which individuals would accept or could not reasonably reject if they were in certain hypothetical counterfactual circumstances. Given that as a consequence these views share the structure of the Basic

Equation above, they too commit the conditional fallacy. The interesting question, however, is whether there is a local way of avoiding the conditional fallacy in this context.[3]

In the next section, I will begin by outlining Nicholas Southwood's recent contractualist theory 'deliberative contractualism' (Southwood 2010). I will argue that it suffers from the conditional fallacy. In Section 3, I will then consider Southwood's responses to my objections (Southwood 2012).

After this, in Section 4, I will turn to a popular standard version of contractualism inspired by Thomas M. Scanlon (Scanlon 1998, chs. 4–5). I will argue that this theory can deal with the problems of Southwood's theory. However, I will also argue that the standard contractualist position too commits the conditional fallacy which becomes apparent in slightly different kind of cases.

In Section 5, I will reformulate Scanlon's contractualism in a slightly new way. In Section 6, I will then show how this proposal works as a local solution to conditional fallacy problem in this context. My solution will also have another advantage as a side-effect. It will provide a new solution to the old problem of at what level of social acceptance we should compare alternative sets of moral principles in Scanlon's contractualism, or so I will argue in the end of this chapter.

II Deliberative Contractualism and the Conditional Fallacy

According to Southwood's deliberative contractualism, 'morality's foundations are to be located in facts about what common code we would agree to live by if we were ... perfectly *deliberatively rational* (Southwood 2010, 86).' As a theory of wrongness, this view can be formulated in terms of the following basic equation (see Southwood 2010, 15):

(Wrongness$_{\text{Southwood}}$) X is wrong if and only if, and just because, if we all were perfectly deliberatively rational, then we would agree to live by a common code which would forbid X.

Let us unpack this thesis further. First of all, Southwood is explicit that he is not offering Wrongness$_{\text{Southwood}}$ as a reductive analysis of the nature of the property of wrongness. Rather, he is offering it as an account of the unique property in virtue of which acts have a further property of being wrong (Southwood 2010, 178). Being forbidden to do some act by the

[3] Previously, the conditional fallacy has most often been discussed in normative ethics in the context of virtue ethics (see Johnson 2003).

common code we would agree to live by if we were perfectly deliberatively rational is, on this view, what makes that act wrong.

Secondly, we need to get some idea of what the relevant hypothetical situation is like in which we are to agree to live by the common code. In those circumstances, we would all be perfectly deliberatively rational.[4] This means that we would, by definition, follow at least three guidelines.

We begin from our actual circumstances in which we live our concrete individual lives. We all have our actual psychological make-ups which consist of our desires, preferences, beliefs, normative judgments, and the like. We then go to a resembling counterfactual situation in which we keep these attitudes constant but in which we are idealised so that we fully comply with the norms of deliberative rationality.

This means first of all that, in those circumstances, before any decision to act, we always deliberate with others who are affected by our decision (Southwood 2010, 89–90). Thus, before any decision to act, one actively engages with others – there is a deliberative 'back-and-forth'. This back-and-forth consists of 'free and open' exchange of relevant information (including the objects of deliberation, the participants' preferences and so on). This communicative element of deliberative rationality includes communicative norms such as the norms of sincerity, intelligibility, and openness (Southwood 2012).

Second, because we satisfy the norms of deliberative rationality in these hypothetical circumstances, our co-deliberation will consist of 'arguing and persuading one another to act in this or that way, while remaining amenable to be persuaded in turn' (Southwood 2010, 90–91). This discursive aspect of deliberation means that we present to each other what we 'take to be considerations for and against options that others are capable of recognising as normatively salient (ibid.).' By doing so, we will attempt to forge a consensus across different perspectives about what we are to do.

Finally, in order to satisfy the norms of deliberative rationality, our co-deliberation must also be reflective (Southwood 2010, 91). We will be 'working out and rendering coherent the content of [our] beliefs and desires, hopes and fears, goals and commitments (ibid.).' We will also be trying to improve ourselves by reorienting the content of our attitudes in the light of our communication and discourse with others. In order to be able to

[4] In addition to this rationality constraint, the hypothetical situation also satisfies inclusiveness and assignment constraints (Southwood 2012). The inclusiveness constraint states that *everyone* is assumed to be involved in the agreement, and the assignment constraint stipulates that we have a specific task of agreeing upon a common code by which to live. The latter consists of both accepting and complying with the relevant principles.

communicate with others freely and frankly, we must know our minds and be prepared to rethink our antecedent attitudes.

Assume then that we go through intensive and prolonged co-deliberation about how we are to live together in a way that complies with the previous norms. At the end of this co-deliberation, we would presumable all agree to live by some common code. According to Southwood, being forbidden to do something by that very code makes acts also wrong in our actual circumstances.

Unfortunately, this appealing view commits the conditional fallacy. The problem is that various acts which are intuitively wrong in the actual world have significantly different qualities in the circumstances in which the norms of deliberative rationality are followed. That in the hypothetical counterfactual situation the norms of deliberative rationality are followed transforms our world, the acts that are done it, and the moral principles that are needed for those circumstances.

Consider the acts resolving disagreements by coercion (see Southwood 2010, 181). Even if these acts are fairly often done in our actual circumstances, no one ever commits them in the relevant hypothetical circumstances in which the norms of deliberative rationality are fully accepted and complied with. If any of these acts were committed, this could only show that we have not yet found the relevant counterfactual scenario we are supposed to consider. Furthermore, given that we would be maximally self-reflective by definition in those circumstances, it would also be common knowledge in the right counterfactual circumstances that no one ever commits these acts.

However, if these acts are never done in the relevant counterfactual circumstances and everyone knows this, our fully deliberatively rational versions would not be motivated to accept a code that made these acts forbidden. Furthermore, they would lack any reason to do so. This is at least for two reasons.

Firstly, the aim of the hypothetical perfectly deliberatively rational group of agents is to construct a set of moral norms that would offer co-operative solutions to their conflicts of interests. However, when it comes to solving disagreements by coercion, there would just be no conflicts or disagreements between the members of this group, and so no moral principles would be needed to solve them. So, given their task, the parties of the relevant hypothetical agreement would not even begin to consider the moral norms that would govern these acts.

Secondly, the parties who are attempting to reach an agreement would know that the principles they choose to govern solving disagreements by coercion would not have any consequences for their behaviour in their circumstances. As long as they remain

deliberatively rational, they will never resort to coercion anyway. This means that the moral principles that would forbid these acts would not have any benefits for the agreeing parties. However, the additional principles would make the set of agreed upon principles more complicated and thus add to the code's internalisation and inculcation costs (Hooker 2000, 78–80). This would provide at least some reason not to include the additional principles in the hypothetical agreement.

Because of these two reasons, I believe that the relevant hypothetical agreement between the perfectly deliberatively rational agents would not include principles that would forbid solving disagreements by coercion. This reveals the conditional fallacy committed by Southwood's theory. According to that view, an act is wrong only if it were forbidden by a code we would agree to live by if we were fully deliberatively rational. Intuitively, it is wrong for us to solve disagreements by coercion. Yet, it is unlikely that we would agree to live by a code which forbade acting in this way if we were fully deliberatively rational. This is because in those circumstances those actions would just not be done and everyone would know this. This means that, because of its subjunctivist structure, $\text{Wrongness}_{\text{Southwood}}$ gets the extension of wrongness wrong.

III Southwood's Responses

In this section, I want to consider the ways in which Nicholas Southwood has tried to argue that his deliberative contractualism can avoid committing the previous conditional fallacy. It is worthwhile to note that, already in *Contractualism & the Foundations of Morality*, Southwood discussed one way in which his view could be reformulated in order to avoid the conditional fallacy (Southwood 2010, 136, fn. 34). This way of avoiding the problem would consist of adopting a version of contractualism inspired by Smith's advice-model. The resulting theory would state:

($\text{Wrongness}_{\text{Southwood}}$*) X is wrong if and only if, and just because, if perfectly deliberatively rational versions of us would co-deliberate, then they would agree that we live by a common code which forbids X in the actual world.

Southwood himself ruled out this theory for good reasons – namely, because it's conceptually incoherent (Southwood 2010, 136, fn. 34). It is impossible for our fully deliberatively rational versions to agree that we, the actual people, live according to some common code.

They can agree what code they are to live by. Only we, the real people, can agree on how we are to live. Because of this, as Southwood was aware when writing his book, Smith's local solution cannot be used to save deliberative contractualism from the conditional fallacy.

After the publication of his book and in response to my objection, Southwood has come up with at least two different new ways of trying to avoid committing the conditional fallacy (Southwood 2012). The first of these responses is a further development of the previous advice-model contractualism. As Southwood has put it, the problem with that proposal is that, in it, our fully rational versions would be making *unconditional* decisions about how we, the actual people, are to live. And, this is something that they just cannot do.

An alternative would be to think that our fully rational counterparts can be understood to be making *conditional* decisions about the principles that govern how they themselves would live in our actual, non-ideal circumstances. Conditional decisions are commonplace in ordinary life. I have, for example, decided to buy a house if I ever happen to buy a lottery ticket and win. Similarly, our deliberatively rational versions could perfectly well decide to agree to live by rules that forbid solving disagreements by coercion in the circumstances in which they are less than perfectly deliberatively rational. As a result, deliberative contractualism could be formulated in the following way (Southwood 2012):

(Wrongness$_{\text{Southwood}}$**) X is wrong if and only if, and just because, if perfectly deliberatively rational versions of us would co-deliberate, then they would agree that (if they are in our actual, non-ideal circumstances, they live by a code that forbids X).

This formulation of contractualism would then presumably avoid committing the version of the conditional fallacy which was described above.

Unfortunately, as Southwood himself has pointed out, this theory too commits the conditional fallacy in certain kinds of cases in which non-compliance is typical in the actual world (Southwood 2012). Consider the question of how much money it is right for our fully deliberatively rational versions to give to charity in their own circumstances. According to Wrongness$_{\text{Southwood}}$**, we must answer this question by considering what kind of rules our fully rational versions would agree to live by in our actual, non-idealised circumstances.

One feature of the actual, non-ideal circumstances is that most affluent people give very little money to charity to alleviate extreme poverty and suffering. Now, Southwood's intuition is that, in this case, the fully deliberatively rational versions of us would follow the

example of the actual affluent people and agree to give very little money to charity in the actual circumstances in which others are doing so too. My intuition is that, in fact, the fully rational versions of ourselves would agree to give vastly more money to charity in these actual circumstances in which there is so much poverty and suffering.

In any case, if we follow Southwood's intuition, then $\text{Wrongness}_{\text{Southwood}}$** entails that the fully rational versions of ourselves should also give very little money to charity in their own circumstances in which everyone follows the norms of deliberative rationality. In this case, not enough money would be donated to help the suffering. Likewise, if we accept my intuitions about the imperfect compliance situation, then it would be wrong for the fully rational versions of ourselves not to give a vast amount of wealth to charity in their own circumstances. In this situation, far too much money would be donated. Thus, both of these results are counterintuitive.

What we really want to say is that the fully rational versions of ourselves would be required to give a moderate amount of their money to charity in their own circumstances in which everyone is doing the same. This amount would be what is required to alleviate the extreme suffering and poverty in their world divided by the amount of fully rational and affluent people living in it. This means that even the revised advice-model contractualism suffers from the conditional fallacy in the situations of imperfect compliance.

Because of this, Southwood has given up the idea of trying to defend his theory by adopting the advice-model. Instead, he believes that the simpler example-model can already avoid my objection if we consider more carefully what moral principles are for (Southwood 2012).

The crux of my argument against the original $\text{Wrongness}_{\text{Southwood}}$ view was that the fully deliberatively rational versions of ourselves would not agree to live by a common code that forbids resolving disagreements by coercion. This is because such agents would already automatically be avoiding these acts and this would furthermore be common knowledge.

In the end, Southwood has decided to respond to my conditional fallacy objection by trying to challenge this core element of my argument (Southwood 2012). That is, he takes on the challenge of explaining why our fully deliberatively rational selves would in fact agree to live by a code that forbids the previous acts even if they are never done.

The essence of his response is the idea that moral principles are not only adopted in order to motivate us to act in a certain way. After all, for this purpose, the principles that

forbid solving disagreements by coercion would be superfluous.[5] Rather, we also adopt moral principles to serve an 'expressive function'. For example, even if no rapes occurred, we would adopt a moral principle that forbids rape in order to 'convey something about the status of women and their right against bodily assault (Southwood 2012).' The idea then is that, in a similar fashion, even if we were fully deliberatively rational, we would agree to live by a code that forbids solving disagreements by coercion because by doing so we could express that coercing others is not all right.

Despite Southwood's own worries, the resulting combination of $\text{Wrongness}_{\text{Southwood}}$ and emphasising the expressive function of moral principles can probably avoid committing the conditional fallacy.[6] My own worry about this proposal, however, is that it makes right and wrong either objectionably relative to our own moral commitments or completely unknowable.

We can recognise this dilemma when we consider the question of who is expressing something by making the fully rational versions of ourselves adopt the principles that forbid solving disagreements by coercion. One answer to this question is that we, the actual people, express our fundamental moral commitments by claiming that the fully deliberatively rational versions of our selves would agree on certain moral principles. But, this would entail that, according to deliberative contractualism, what is right and wrong depends in some fundamental sense on our basic moral commitments. This is the way to objectionable relativism. If our basic moral commitments had been different, then perhaps solving disagreements by coercion would not have been wrong after all.

To avoid this consequence, Southwood could claim that it is the fully rational versions of ourselves themselves that express their fundamental moral commitments by agreeing to live by certain moral principles. But, then, what motivation is there to think that the fully idealised versions of ourselves would find it necessary to express something by agreeing to live by principles that forbid solving disagreements by coercion? The answer to this question cannot merely be that in this way the theory would avoid committing the conditional fallacy as this would be objectionably *ad hoc*.

[5] Southwood claims that the principles could still a play a role in motivating the fully rational agents to keep conforming to the norms of deliberative rationality in the future (Southwood 2012). But, it's not clear why the fully rational agents would not then adopt conditional principles for the future situations in which they no longer are fully rational. This, however, gets back to the conditional fallacy committed by $\text{Wrongness}_{\text{Southwood}}$**.

[6] Southwood is worried about requirements to do the second best. If you are angry at someone for no good reason at all, then you are required to apologise. You have this duty even if your first duty is not to be angry in the first place. Southwood worries that the example-version of deliberative contractualism could not explain this duty. I assume that he could claim that the fully rational agents would accept the duty to do the second best for expressive reasons too.

As a result, we would need to know much more about what the fully deliberatively rational versions of ourselves would want to express by adopting different moral principles. However, short of relying on our own fundamental moral commitments, I see no other way of knowing what moral commitments they would want to express and why. In this way, Southwood's reliance on the expressive function of moral principles makes his view epistemically too demanding. It also creates a problem for his thesis that the hypothetical agreement can on its own play a foundational role in explaining what is right and wrong (see Southwood 2010, sec. 1.2). After all, now we are relying on our own fundamental moral commitments in order to determine what the fully rational versions of ourselves would express by agreeing to live by certain principles. Because of these problems, I want to move onto consider whether other forms of contractualism also commit the conditional fallacy and whether they could avoid it.

IV Standard Contractualism and the Conditional Fallacy

In this section, I will first introduce a standard version of contractualism which is very much inspired by Scanlon's view in his *What We Owe to Each Other* (Scanlon 1998). I will then argue that this theory avoids the version of the conditional fallacy which is committed by Southwood's view. Finally, I will construct a case which illustrates how the standard version of contractualism too commits the fallacy.

According to Scanlon, 'an act is wrong if and only if... it would be disallowed by any principle that ... people could not reasonably reject (Scanlon 1998, 4).' The first thing to notice about this view is that the notion of agreement, either actual or hypothetical, plays no role in it (Scanlon 2004, 125, 134).[7] This is why the theory does not guide us to consider which moral principles would be agreed upon in any counterfactual circumstances. Rather, we must consider which set of moral principles could not be reasonably rejected by anyone.

Here is a sketch of how I will understand the notion of reasonable rejectability. We begin from a set of possible worlds which are otherwise exactly like our world but different sets of moral principles have been adopted in them. In fact, there is a possible world for every potential set of principles which we could adopt such that that very code has been internalised in it.

[7] It is worthwhile to note that, because of this, the standard form of contractualism is not a contractualist theory by Southwood's lights (Southwood 2010, 3).

Whichever moral code has been internalised in a given world significantly affects which actions are done in that world. In a world whose code forbids lying, few lies are told.[8] Furthermore, which actions are done in a world has a significant impact on what kind of lives people come to live in that world. Our lives can go better or worse because of how we and others act. Let us call the kind of life which an individual lives under a given set of moral principles her 'standpoint' (Scanlon 1998, sec. 5.4).

Some standpoints are more choiceworthy than others. Some elements of the standpoints make them lives which we would wish to live, whereas other elements of those standpoints make them lives which we would want to avoid. Let us then call the features of the standpoints that make them less choiceworthy 'burdens'.

Whether a given set of moral principles is then reasonably rejectable is a function of the burdens which individuals need to bear under that and the other alternative codes. When a moral code makes someone's life burdensome, that individual can make an objection to that moral code. The more serious the personal burden caused by the code is, the more serious personal objection the individual who has to bear that burden can make.

In this situation, an individual can reasonably reject her world's moral code which is creating serious burdens for her whenever there is an alternative code to which no-one can make equally serious personal objections on the basis of their burdens. This means that the non-rejectable moral code is such that there are more serious personal objections to all other moral codes. All those codes produce more burdensome lives to some individuals. According to the standard version of contractualism, an act in the actual world is then wrong if and only if it is forbidden by the non-rejectable code understood in the way explained above. The standard version of contractualism could thus be captured by the following kind of Basic Equation:

($\text{Wrongness}_{\text{standard}}$) X is wrong if and only if, and just because, if different moral codes were adopted in different possible worlds, then X would be forbidden by the moral code such that there would be more serious personal objections in these worlds to all other codes.

The question then is: does this theory also commit the conditional fallacy? First of all, it does not suffer from the problems of Southwood's theory faces. The standard version

[8] We need not assume that acceptance entails perfect compliance because of weakness of will and other irrationalities (Hooker 2000, sec. 3.2).

of contractualism can easily explain why resolving disagreements by coercion is wrong. We begin by comparing two possible worlds. In one world, everyone has adopted a moral code which authorises these actions, whereas the people of the other world have adopted a code which forbids them. We then look at what kind of lives people come to live in these two worlds as a result of these actions being done in one of the worlds but not in the other.

In the world in which certain acts of coercion are considered to be permissible and therefore occasionally done, the victims will have to bear serious burdens as a consequence. Because no-one needs to bear similar burdens in the other world, we can conclude that these acts are forbidden by the principles which no-one could reasonably reject. As a consequence, according to the standard contractualist view, these acts are wrong.

This result might make us think that the standard version of contractualism avoids committing the conditional fallacy. There might also be structural reasons to draw this very conclusion. The subjunctive conditional on the right-hand side of $\text{Wrongness}_{\text{standard}}$ does not guide us to consider what happens to X in certain counterfactual circumstances. Rather, it makes us to compare how X behaves under a vast number of different counterfactual situations. One might hope that this avoids changing the nature of X in the comparisons so as to make the consequences irrelevant for whether X is right or wrong in the real circumstances.

Unfortunately this optimism is not warranted. There are familiar cases which pose the conditional fallacy problem even for the standard version of contractualism.[9] Imagine that Rick is a person who lies, cheats and steals. Because of Rick's actions, many people come to suffer. However, Rick is not rotten to the core but rather he wants to become a better person. All he needs is someone to teach him how to be a bit more virtuous. One day, Rick sees a poster of a course that promises to make him more virtuous. Is Rick required to go on the advertised course?

According to the standard version of contractualism, the answer to this question depends on what is required by the moral code which is internalised in the world in which individuals have to bear the least serious personal burdens. The question is, does the moral code internalised in that world require individuals like Rick to go on a course that teaches virtue?

To answer this question, we must compare two worlds. In one world, the internalised code does not require bad people to go on courses that teach the basics of virtue. This

[9] This argument follows Hooker's rule-consequentialist discussion of similar material (Hooker 2000, sec. 3.3). The case is modified from Johnson (2003, 816–818).

element of the code has no consequences to anyone's lives in that world because the internalised moral code already makes everyone sufficiently virtuous in that world. In the other world, the moral code which is internalised by everyone is almost exactly like the previous code with the exception that it, in addition, requires individuals like Rick to go on the virtue courses. Given that, in this world too, everyone has by stipulation adopted the relevant moral code, the requirement to go on those courses does not apply to anyone in this world either. Everyone is already sufficiently virtuous in this world too. So, it might initially seem like both of these codes create the same standpoints to all individuals, and thus neither one of them could be reasonably rejected.

This is problematic for the standard version of contractualism. On one interpretation of the theory, given that one non-rejectable set of principle does not require Rick to go on the course and another one does, Rick's act would be both required and not required. This would mean that the theory has contradictory consequences. An alternative would be to say that, if an act is both required and not required by different non-rejectable sets of principles, then it is merely permissible for Rick to go on the course. However, this view would fail to capture the intuition that Rick really ought to go on the course.

Furthermore, the code that does not require anyone to go on the virtue courses in one of these worlds could help the individuals of that world to avoid small personal burdens. That code would be simpler than the alternative code with the additional requirement that never applies. It would contain one prescription less to internalise. This might make the first code just a little bit less burdensome to learn (Hooker 2000, 78). This would mean that it could not be reasonably rejected whereas the alternative could be. This too would make it the case that Rick is not required to go on the course he needs in our world.

The point of all of this is to illustrate how the standard version of contractualism too commits the conditional fallacy. The antecedent of its counterfactual conditional assumes that different moral codes have been adopted by everyone in the compared worlds. This means that there is a class of acts that no one ever needs to consider in these worlds. These are the acts of self-improvement that would be required for internalising the moral codes of those worlds. For this reason, comparing the personal burdens in these worlds can tells us little about the moral status of such acts in our world in which this type of acts are sometimes required. This problem too has the structure of the conditional fallacy.

V Reformulating Contractualism

In order to avoid the previous problem, I believe that we should reject one implicit assumption of the standard version of contractualism.[10] This assumption is the stipulation that *everyone* has adopted the relevant moral codes in the compared worlds. This stipulation is not realistic, and it serves no useful purpose.

Rather, in the following, I will assume that human agents have only an innate capacity to internalise any one of the potential moral codes. In order to make use of this capacity, the children who will adopt the moral codes will need to be brought up: taught and educated. They will also need to interact with others in the moral community – take part in the moral practices which includes reacting to the actions of others in different ways and correcting these reactions. And, they need to practice and come across admirable moral exemplars, significant moral problems, and difficult moral dilemmas. Only through this process can individuals come to internalise a moral code.[11]

I will also assume that how natural or burdensome it is to adopt a moral code depends on the content of the moral code to be adopted and the individuals who will do the adopting. Perhaps ordinary human beings have biological dispositions to adopt some moral codes more easily than others. Likewise, maybe some individuals are more morally malleable than others. Just as some people learn mathematics easily whereas this is very hard work for others, it could be that some individuals learn morals naturally whereas, for others, this is a hard and difficult process.

With this in mind, I propose that we should understand the moral codes which are compared in the Scanlonian framework to consist of two separate elements. The first element contains the principles that describe how the mature agents who have internalised that moral

---

[10] So far I have used the slightly awkward name 'the standard version of contractualism inspired by Scanlon' for the theory under discussion. A simpler name for this view would just be 'Scanlon's contractualism'. However, the problem is that this name would mean that, when Scanlon described his version of contractualism, his idea was that everyone in the relevant worlds have adopted the evaluated sets of moral principles. Scanlon himself only wrote that:

> According to contractualism, in order to decide whether it would be wrong to do X in circumstances X, we should consider possible principles governing how one may act in such situations, and ask whether any principle that permitted one to do X in those circumstances could, for that reason, reasonably be rejected. In order to decide whether this is so, we need first to form an idea of the burdens that would be imposed on some people in such a situation if *others* were permitted to do X (Scanlon 1998, 195, my emphasis).

I have always read 'others' in this passage to stand for 'everyone'. However, as an anonymous referee has pointed out to me, this might not have been Scanlon's intention and it is true that Scanlon never discusses the problems that follow from assuming the 100% internalisation rate. Because of this, I do not want to claim that Scanlon committed himself to the standard version of contractualism I have described above. This explains why I am using the slightly awkward name for the discussed view.

[11] It is better to understand internalising a moral code as establishing a moral conscience or sensitivity rather than as learning to remember a set of principles (Hooker 2000, sec. 3.5).

code are to act. These are the familiar standard ethical principles. Call this the 'basic element' of a set of moral principles.

The second part of the moral code describes (i) which actions mature moral agents are required to do in order to get the next generations to internalise their code, (ii) what additional corrective requirements there are for those agents who fail to internalise the moral code as children, and (iii) what the moral community can demand from those agents. Call this the 'inculcation element' of a set of moral principles.

Here is then the basic idea which will ground my proposal: what percentage of a new generation will internalise a given moral code depends on how much effort the previous generation puts into the moral upbringing of that new generation.[12] To illustrate this thought, let us take a moral code whose basic element is much like our ordinary common-sense morality. It forbids killing, stealing, cheating, breaking contracts, lying, hurting others, failing to satisfy the urgent needs of others, and so on. It also offers some guidance to what to do in the conflicts of these principles, and what exceptions they have.

We can attach different inculcation elements to this moral code. Some of them require intensive hands-on parenting and thorough moral education in social institutions such as schools. Other more minimal inculcation elements do not require any acts of moral education from the parents or the wider social circle within which the children grow up. And, of course, there is a whole spectrum of different kinds of potential inculcation elements of codes between these two extremes.

The next idea is that how many individuals of every new generation in a world come to internalise the basic element of a code depends on which inculcation element of the code has been adopted in that world. We can take a set of worlds in which the basic element of the moral code in each world is the same whereas every world has a different inculcation element attached to that basic element. So, in some worlds, the adopted code requires putting a lot of effort in getting the next generations to internalise the code, whereas in other worlds the minimal inculcation element fails to guide individuals to transfer their codes to the next generations.

As a result, because of the different inculcation elements of the moral code, the level of social acceptance of the given basic elements of the code will be different in these worlds. Some very intensive inculcation elements might eventually get the level of acceptance in the

[12] In the rule-consequentialist framework, this point was first made by Holly Smith (2010, 416).

world close to 100%. Some very minimal inculcation elements might not get that level much above 0% in the end.

Here I assume that, for every combination of a basic element and an inculcation element as a set of moral principles, there is a stable and self-sustaining equilibrium state such that the same percentage of each new generation comes to internalise that particular set of moral principles.[13] This is because how many percent of a new generation comes to internalise the given code is a function of how many people in the previous generation had internalised the code and how much effort they will put in inculcation given the code they have internalised.

How high the percentage of internalisation is at the equilibrium point will thus depend on how demanding the inculcation element is. I assume that, if we set a number of worlds with a given combination of a basic element and an inculcation element in motion from a set of fairly high percentages of internalisation (perhaps ranging from 100% to 60%), these worlds will eventually converge upon that code's own equilibrium state. For codes with undemanding inculcation elements the equilibrium point can of course be at 0% of internalisation, but for many other codes that point will presumably be much higher.

We can then compare consequences of the different combinations of basic elements and inculcation elements at their equilibrium points. Each of these combinations will create different standpoints for the individuals in the worlds in which they have been adopted. We then compare the personal objections to these more extensive combinations of the two elements of moral codes, and choose the code to which individuals have the least serious personal objections.

Here is a rough estimation of what would happen. The more a world puts effort into inculcation of the next generations, the higher the level of social acceptance of the code will be. This leads to fewer acts that create unnecessary personal burdens – killings, stealings, cheatings, promise-breakings, and so on.[14] However, the more effort is used to inculcate, the

[13] Some combinations of basic elements and inculculation elements can have many equilibrium points at different level of social acceptance. In this situation, we should consider the equilibrium point at which the lives of individuals are the least burdensome (see Smith 2010, 417–418). Other combinations might not have a set equilibrium but rather their social acceptance might go up and down in a pattern in the same way as the size of animal populations often varies. We can compare these combinations to other codes by considering the burdensomeness of individual lives throughout the cycle.

[14] Here I am assuming that the more effective an inculcation element is, the more it will require effort and the more burdensome it will therefore be. As an anonymous referee correctly observed, this may not always be true. As he or she nicely puts it, 'depending on the code, and the ingenuity involved in the inculcation plan, it seems that in principle we might have cases where we have highly effective inculcation element but with very little effort or burden.' This seems true to me. It only becomes a problem for the position described below if the least burdensome combination of a basic element and an inculcation code has its equilibrium point at 100%

more burdensome this will become for many individuals. This is not only due to the sheer time and energy required for the intensive education, but also due to the fact that, at some point, intensive moral education begins to clash with freedom of conscience, freedom of speech, and autonomy. And, not having these freedoms would be a significant burden for individuals.

This means that, in a world in which both elements of a moral code together lead to the least serious personal objections to that code, the efforts in moral education are unlikely to produce universal acceptance of the code in the new generations (even if they might produce a high rate of acceptance). To get 100% acceptance rate would just be too burdensome to some individuals when compared to the personal burdens caused by the actions of those few individuals whom the society failed to inculcate. The set of principles that could not be reasonably rejected would thus not be accepted by everyone because getting the remaining people to accept the code would create more serious burdens.

VI Advantages of the New Proposal

I want to then finish off by explaining two advantages of this proposal.[15] Firstly, it avoids the conditional fallacy committed by Scanlon's view. Scanlon's view committed that fallacy because it assumed that the alternative moral codes are universally accepted in the compared worlds. For this reason, that view could not explain why a bad agent would be obliged to develop his character.

My proposal avoids this problem. This is because, according to it, the compared codes will include an inculcation element, and so also the non-rejectable code will include some principles of inculcation. In a world in which agents were never required to develop

---

acceptance rate. If absolutely everyone could be captured with an inculcation element which required little effort and created few burdens, my view will not be able to deal with cases like Rick in the real world. I believe that the modest assumption that such a perfect inculcation element does not exist is reasonable.

[15] I acknowledge that my proposal also faces serious problems, which unfortunately I do not have space to explore here. The first class of problems are general problems of contractualism such as the aggregation problems (Suikkanen 2004), redundancy problems (Suikkanen 2005) and the problem of duties towards non-humans (Hooker 2000, 66–70). Of course all forms of contractualism will have to deal with these issues in one way or another. The second class of problems are special problems for all versions of contractualism that do not assume 100% acceptance of the compared sets of principles. An especially difficult challenge is the question of what the individuals who have not internalised the compared codes should be assumed to be doing as this will have significant consequences for how the codes will affect individuals (see Smith 2010, sec. 5). I briefly sketch the alternative solutions to this problem in Suikkanen (2013). It seems to me that we should bear the costs of these solutions rather than accept a 100% acceptance version of contractualism that cannot account for duties of moral education, reparation and gratitude. Finally, it might be that there are special objections to my theory that do not apply to the other versions of contractualism that do not require 100% acceptance of the compared codes (see below). I will leave it to my critics to investigate what such problems could be.

their characters as mature agents, fewer people would have internalised the moral code of that world. This would entail that, in that world, more actions which cause personal burdens would be done. And, so those principles could be reasonably rejected because there is an alternative that does not cause as serious burdens to anyone.

The alternative set of principles that leads to less burdensome lives includes some principles of what agents like Rick are required to do in order to improve their characters. These non-rejectable principles also include a principle that requires others to guide Rick to make correct choices about how to develop his character. So, it does seem to me that my proposal can avoid the conditional fallacy which is committed by both Southwood and Scanlon.

This proposal has also another advantage. In the discussions about rule-consequentialism, it is widely accepted that it is a bad idea to compare the consequences of 100% acceptance of the alternative moral codes.[16] For one, if we compared the alternative sets of principles, as explained above, we could not explain why characters like Rick can be required to improve their character. As result, various alternatives have been proposed. They could also be applied in the contractualist framework. I want to finish off by explaining the advantages which my proposal has over these alternatives.

The first alternative is just to pick a number. So, according to Brad Hooker, we should compare the consequences of 90% of people in each new generation internalising the compared codes (Hooker 2000, 84). We could then consider which code could not be reasonably rejected if it were accepted in each new generation by 90% of people. This comparison would presumably create some duties which govern how we should treat those remaining 10% of the population who have not accepted the relevant codes by stipulation. Lacking such duties would make the life of the potential victims of those individuals more burdensome. So, this alternative could create a duty for Rick to improve his character.

This proposal faces two challenges. Firstly, the number 90% is obviously arbitrary and this seems theoretically unattractive.[17] It could just as well be 85% or 92%. Secondly and more seriously, there are two ways in which 90% of the new generations can come to adopt a given code. We can either just stipulate that 90% innately accept the alternative codes in the different worlds, or we can think that the previous generations have a duty to inculcate 90% of the new generations. The former option would fail to generate duties for us

[16] See, for instance, Hooker (2000, sec. 3.3).
[17] See Ridge (2006, 246-248); for a response, see Hooker & Fletcher (2008).

in the real world to educate our children morally. In contrast, the latter alternative would generate some moral duties to inculcate the future generation for us.

However, the problem is that in this situation these duties would lack a rule-consequentialist or a contractualist justification. We would not have these duties because either having them would have the best consequences or because these duties could not be reasonably rejected. Instead, these duties would be ascribed to us solely because they give us the magical 90% internalisation rate. But, why would this give us any reason to comply with these inculcation principles?

The second alternative is to think of the consequences of the acceptance of the codes at each level of social acceptance (at 100%, and at 99%, and ..., and at 0%) (Parfit 2011, 317-318). We then consider what kind of objections people will have to the alternative codes at each level of acceptance. The non-rejectable code would then be the one to which there is least serious personal objections at every level of social acceptance.

The problem with this proposal is that it is unlikely that there is any one code such that it creates the least serious personal burdens at every level of acceptance (see Ridge 2009, 67). It can well be that one code would not lead to any serious objections if were accepted by almost everyone, whereas that same code could lead to serious avoidable objections when only few people have adopted it. In this situation, this second proposal would mean that no set of principles would be non-rejectable and thus no action would be wrong.

Michael Ridge has suggested a way to avoid this problem (Ridge 2006). He too believes that we should consider the consequences of moral codes at each level of social acceptance. Each level of acceptance creates different lives for the individuals of the compared worlds. This means that, for each individual, there is an 'average burdensomeness' of all the lives which that individual would live under the different levels of social acceptance of the given code. We could then consider to which set of principles individuals could present the least serious personal objections on the basis of the average burdensomeness of their lives. This would be the non-rejectable code according to the 'variable-rate contractualism' which I have formulated on the basis of Ridge's 'variable-rate rule-utilitarianism'.

This proposal seems to suffer from a problem, which my proposal explained above can avoid.[18] The compared sets of principles would need to include some principles that govern which actions individuals are required to do to morally educate the next generations.

[18] For further problems, see Hooker & Fletcher (2008, sec. 4).

However, it's not clear how we could assess, in Ridge's framework, the consequences of different potential inculcation elements. We would have to consider these elements too on different levels of social acceptance. So, we would have to take, for instance, a very demanding inculcation element and consider its consequences over time on 95% level of acceptance, 50% level of acceptance, and 10% acceptance. However, we could not seem to be able to do this because the level of acceptance itself depends on the demandingness of the compared inculcation element. It's not clear how you could consider a demanding inculcation element on a low level of acceptance because the inculcation element itself would raise that level immediately.

This shows that Ridge's proposal could not be used in the contractualist framework to generate the moral principles that govern moral education. However, my proposal above was able to do so. It leaves room for taking into account how alternative inculcation elements of moral codes can affect the level of social acceptance of codes, and it provides a contractualist way to evaluate which of these alternatives then belongs to the non-rejectable set.

Bibliography

Bird, Alexander (1998): "Dispositions and Antidotes". *Philosophical Quarterly* 48: 227–234.

Bonevac, Daniel, Dever, Josh & Sosa, David (2006): "Conditional Fallacy". *Philosophical Review* 115: 273–316.

Fara, Michael (2005): "Dispositions and Habituals". *Noûs* 39: 43–82.

Harman, Gilbert (1975): "Moral Relativism Defended". *The Philosophical Review* 84: 3–22.

Hooker, Brad (2000): *Ideal Code, Real World* (Oxford: Oxford University Press).

Hooker, Brad & Fletcher, Guy (2008): "Variable versus Fixed-Rate Rule-Utilitarianism". *Philosophical Quarterly* 58: 344–352.

Johnson, Robert (1999): "Internal Reasons and the Conditional Fallacy". *Philosophical Quarterly* 49: 53–71.

Johnson, Robert (2003): "Virtue and Right". *Ethics* 113: 810–834.

Johnston, Mark (1992): "How to Speak of the Colors". *Philosophical Studies* 68: 221–263.

Johnston, Mark (1993): "Objectivity Reconfigured: Pragmatism and Modern Idealism". In C. Wright & M. Haldane (eds.): *Reality, Representation, and Projection* (New York: Oxford University Press), 85–130.

Lewis, David (1997): "Finkish Dispositions". *Philosophical Quarterly* 47: 143–158.

Martin, C.B. (1994): "Dispositions and Conditionals". *Philosophical Quarterly* 44: 1–8.

Parfit, Derek (2011): *On What Matters, Vol 1* (Oxford: Oxford University Press).

Ridge, Michael (2006): "Introducing Variable-Rate Rule-Utilitarianism". *Philosophical Quarterly* 56: 242-256.

Ridge, Michael (2009): "Climb Every Mountain?". *Ratio* 22: 55–77.

Scanlon, T.M. (1998): *What We Owe to Each Other* (Cambridge, MA: Harvard University Press).

Shope, Robert (1978): "Conditional Fallacy in Contemporary Philosophy". *Journal of Philosophy* 75: 397–413.

Smith, Holly (2010): "Measuring the Consequences of Rules". *Utilitas* 22: 413–433.

Smith, Michael (1995): "Internal Reasons". *Philosophy and Phenomenological Research* 55: 109–131.

Southwood, Nicholas (2010): *Contractualism & the Foundations of Morality* (Oxford: Oxford University Press).

Southwood, Nicholas (2012): "Does Deliberative Contractualism Commit the Conditional Fallacy?". Presentation at the Future of Contractualism Conference, University of Rennes, France, 12.5.2012.

Suikkanen, Jussi (2004): "What We Owe to Many". *Social Theory and Practice* 30: 485–506.

Suikkanen, Jussi (2005): "Contractualist Replies to the Redundancy Objections". *Theoria* 71: 38–58.

Suikkanen, Jussi (2013): "Rule-Based Ethical Theories and Counter-Cultures". *Pea Soup*. http://peasoup.typepad.com/peasoup/2013/02/rule-based-ethical-theories-and-counter-cultures.html (accessed 8. 3. 2013).

Watson, Gary (1975): "Free Agency". *Journal of Philosophy* 72: 205–220.

Wright, Crispin (1992): *Truth and Objectivity* (Cambridge, MA: Harvard University Press).