Forthcoming in *Philosophy and Phenomenological Research*

CONTRACTUALISM FOR US AS WE ARE[[1]](#footnote-1)

Nicholas Southwood

Australian National University

Abstract: A difficult problem for contractualists is how to provide an interpretation of the contractual situation that is both subject to appropriately stringent constraints and yet also appropriately sensitive to certain features of us as we actually are. My suggestion is that we should embrace a model of contractualism that is structurally analogous to the “advice model” of the ideal observer theory famously proposed by Michael Smith (1994; 1995). An advice model of contractualism is appealing since it promises to deliver a straightforward solution to the so-called “conditional fallacy.” But it faces some formidable challenges. On the face of it, it seems to be straightforwardly conceptually incoherent. And it seems to deliver a solution to the conditional fallacy at the cost of being vulnerable to what I shall call “the concessional fallacy.” I shall consider how, if at all, these challenges are to be met. I shall then conclude by considering what this might mean for the so-called “ideal/non-ideal theory” issue.

I. Introduction

I shall take contractualism to be the view that morality or some part of it (justice, social justice, what we owe to each other) is somehow dependent upon some special, privileged agreement within some special choice situation.[[2]](#footnote-2) Some contractualists hold that the agreements are *actual* agreements: agreements that are supposed to actually happen.[[3]](#footnote-3) This is a non-starter. To avoid making morality arbitrary and internally inconsistent, any version of contractualism must involve stringent constraints on the relevant morality-grounding agreements: inclusiveness constraints; rationality constraints; informational constraints; motivational constraints; and so on. But there are no actual agreements that satisfy these stringent constraints. So, the only plausible versions of actual agreement contractualism entail that there are no actual agreements of the kind that morality is supposed to be dependent on and, hence, no (non-trivial) moral truths.[[4]](#footnote-4)

It shouldn’t be surprising, then, that most contractualists hold instead that the relevant agreements are highly idealised *counterfactual* agreements: agreements that would or could or should happen if we were to satisfy relevant stringent constraints. Counterfactual versions of contractualism are clearly well placed to provide a specification of the contractual situation that is subject to stringent constraints of the kind that is needed to avoid making morality arbitrary and internally inconsistent, while staving off the spectre of moral nihilism. But they face other problems. The problem that I want to focus on here is that it is notoriously difficult to provide an interpretation of the counterfactual contractual situation that is appropriately sensitive to features of us as we actually are.

An important example of this kind of concern is what, in a classic paper, Robert K. Shope (1978) called the *conditional fallacy*. A counterfactual theory commits the conditional fallacy when certain respects in which the relevant counterfactual circumstances differ from the actual world make the relevant counterfactual theory give wrong verdicts in the actual world. This can be illustrated by considering the following counterfactual analysis of ability (CA):

(CA) An agent A is able to X iff A would succeed in Xing if A were to choose to X (Moore 1912).

The problem is that it seems that we are sometimes unable to perform certain acts precisely because there is some insurmountable impediment to choosing to perform the acts in the actual world. To borrow a much-discussed example from Susan Wolf, suppose that “a person attacked on a dark street would have screamed if she had chosen … [but is] too paralyzed by fear to consider, much less choose, to scream” (Wolf 1990, p. 99). It might seem obvious that the victim is unable to scream given the fact that her actual circumstances include being gripped by debilitating terror. (CA) commits the conditional fallacy because it gives the wrong answer about what the victim is able to do in her actual circumstances since the relevant counterfactual circumstances are ones where she has somehow overcome this terror.

It is not hard to see why counterfactual contractualist theories might be vulnerable to the conditional fallacy.[[5]](#footnote-5) The stringent constraints needed to avoid making morality arbitrary and inconsistent mean that there will be significant differences between the superior characteristics and circumstances of the contractors whose agreements morality is supposed to be dependent on and our radically inferior characteristics and circumstances. This opens up the possibility of a salient *mismatch* between the sorts of principles that it would make sense for the contractors to agree to live by and the sorts of principles that would make sense for us as we actually are. Given that contractualists hold that the moral status of *our* actions depends on whether they accord with principles to which the contractors would agree, contractualism will result in both false positives and false negatives. It will result in false positives since it will imply that we are morally permitted (indeed required) to perform acts that are morally impermissible for us to perform given our actual characteristics and circumstances (though they may well be permitted or required of our contractor counterparts given their characteristics and circumstances). And it will result in false negatives inasmuch as it will fail to explain other acts that are morally permissible or obligatory for us given our actual characteristics and circumstances (even though they may not be permissible or obligatory for our contractor counterparts).

My aim in what follows is to consider one appealing way in which a contractualist might seek to accommodate both stringent constraints within the contractual situation and appropriate sensitivity to us as we actually are. This is to embrace a model of contractualism that is structurally analogous to the “advice model” of the ideal observer theory famously proposed by Michael Smith (1994; 1995). An advice model of contractualism is appealing since it promises to deliver a straightforward solution to the conditional fallacy. But it faces some formidable challenges. On the face of it, it seems to be straightforwardly conceptually incoherent. And it seems to deliver a solution to the conditional fallacy at the cost of being vulnerable to what I shall call “the concessional fallacy.” I shall consider how, if at all, these challenges are to be met. I shall then conclude with some rather speculative remarks about what this might mean for the so-called “ideal/non-ideal theory” issue.

II. An illustration: deliberative contractualism and the conditional fallacy

Let’s start with the conditional fallacy. It will be helpful to provide a concrete illustration of how the conditional fallacy might be thought to arise in the context of a particular contractualist theory. The contractualist theory I shall focus on here is the so-called “deliberative” version of contractualism that I have developed and defended elsewhere (Southwood 2010; cf. Habermas 1984-1987; Forst 2014). Deliberative contractualism is a good test case for our purposes since it seems especially vulnerable to the conditional fallacy (Suikkanen 2014, pp. 118-26). If we can find a solution that works in the case of deliberative contractualism, there is every reason to believe that the solution (or something like it) might work elsewhere.

We can formulate deliberative contractualism roughly as follows:

(DC) It is morally impermissible/obligatory for A to X iff if we were all perfectly deliberatively rational and charged with the task of agreeing upon a common code by which to live, then we would agree to live by a common code, C, that includes a principle P that forbids/requires A to X.

The counterfactual agreements upon which deliberative contractualism is based involve three main kinds of constraints. First, the agreements are subject to an *inclusiveness constraint*. We are *all* taken to be included in the agreements. It is natural to ask exactly who is included in this “we.” Let’s say that it involves at least all those deliberatively competent individuals who actually exist. There are interesting and important questions about what to say about a) the deliberatively incompetent, b) deliberatively competent individuals who are either dead or who haven’t been born yet, and c) merely possible deliberatively competent individuals. These questions lie beyond the scope of the current paper.

Second, the agreements are subject to a kind of *assignment constraint*. We are taken to be charged with a specific and rather unusual task, the task of agreeing upon a common code by which to live. A common code by which to live involves a set of shared principles that permit, forbid and require us to do things. Living by a common code involves complying with the principles. I assume it also involves internalising the principles; holding others (and oneself) to account with respect to the principles; perhaps even encouraging others to comply with and internalise the principles; and so on (Hooker 2000).

Third, and most importantly, the agreements are subject to a specific kind of *rationality constraint*. In particular, the agreements are such that we are taken to be *perfectly deliberatively rational*. What does it mean to be perfectly deliberatively rational? It means that we engage in, and make decisions on the basis of, prolonged deliberation with others that fully accords with three kinds of deliberative norms. The first are communicative norms, which are oriented to mutual understanding, regulate the ways in which we exchange information. They include norms of sincerity, intelligibility and openness. The second are discursive norms, which are oriented to mutual persuasion and consensus, regulate the ways in which we argue with one another. They include norms of persuasion, reciprocity and adaptiveness. The third are reflective norms, which are oriented towards self-interrogation and self-transformation, regulate the organisation of our mental states. They include norms of coherence, introspection, and adaptability. To satisfy the deliberative rationality constraint, then, requires that we deliberate in a manner that fully complies with all relevant communicative, discursive, and reflective norms, and that the decisions we reach are based on the outcome of this deliberation.

Deliberative contractualism, then, holds that the moral status of our actions depends on whether or not the actions accord with principles that are the product of counterfactual agreements that satisfy each of these three stringent constraints. A view of this kind might seem especially vulnerable to the conditional fallacy. That’s because it seems very likely that there will be a mismatch between the sorts of principles that would make sense for deliberative contractors and the sorts of principles that would make sense for us as we are.

Jussi Suikkanen gives one possible illustration of how such a mismatch might arise:

The problem is that various acts which are intuitively wrong in the actual world would have significantly different qualities in the circumstances in which the norms of deliberative rationality are followed. … Consider the acts of resolving disagreements by coercion, threats, bribery, blackmail, and so on. 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 complied with. … 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, our fully deliberative rational versions would not have any reason to accept a code that made these acts forbidden. After all, we would all know that no one ever does any of these things. … This reveals the conditional fallacy of Southwood’s theory. … Intuitively, it is wrong for us to solve disagreements by coercion, threats, bribery, blackmail, and so on. Yet it is unlikely that we would agree to live by a code which forbade acting in these ways if we were fully deliberatively rational (Suikkanen 2014, pp. 120-21).

This particular argument strikes me as questionable. For it is not obvious why we should think that if we were such that we never resolve disagreements by coercion, threats, bribery, and blackmail, and that this were a matter of common knowledge among us, then we would not agree to live by a common code that includes principles that forbid resolving disagreements by coercion, threats, bribery, and blackmail. Suikkanen doesn’t say much to justify this claim. But the key seems to be that we only “need” a principle requiring us not to X if some of us sometimes are going to X. This is suggested by his remark that “the counterfactual situation [in which] the norms of deliberative rationality are followed transforms … the moral principles that are *needed* for those circumstances’” (Suikkanen 2014, p. 120: italics added). His thought seems to be that the sole function of principles is a *motivational* function: to encourage us to refrain from behaving in ways in which some of us would be likely to behave if not for the principle. Where we are all already refraining from behaving in those ways, we don’t need a principle forbidding us from behaving in that way; the principle is motivationally superfluous. If the sole function of principles is their motivational function, then the principle is superfluous simpliciter. If the principle is superfluous, then we wouldn’t agree to live by such a principle.

This is too quick. First, it is not obvious that where we are all already refraining from Xing, there is no potential motivational role that a principle forbidding us from Xing might serve. While it may be the case that we are all refraining from Xing *now*, there is presumably some non zero probability that some of us will *change* in the future such that some of us will be tempted sometimes to X;[[6]](#footnote-6) and a principle forbidding us from Xing may make it *less likely* that we change in this way. I am not saying that such a principle makes it more likely, if we do change, that we will refrain from Xing despite being tempted to X. We are concerned with the motivational function that principles serve in circumstances where we are all refraining from Xing, and no one is tempted to X. So it is irrelevant to point to the motivational function that a principle forbidding us from Xing would have in circumstances where some people *are* tempted to X. Rather, the thought is that the principle may serve the motivational function, in our circumstances where no one is Xing or tempted to X, of making it less likely that we change such that some of us are sometimes tempted to X.

Second, it is not obvious that the sole function of principles is their motivational function. Principles might also be thought to serve an *expressive* function. Many scholars have emphasised the expressive function of law or *legal* principles (Anderson and Pildes 2000). But non-legally enforced principles also have an important expressive dimension (Brennan et al 2013, ch. 2). They express what is of importance to us, what we value*.* Take a principle forbidding torture. Clearly this is important in part—perhaps primarily—because of its motivational effects; it makes it less likely that torture will occur. But even we were to occupy a context in which torture was very unlikely to occur, a principle forbidding torture might remain important because of what it conveys about the status of persons. It says, in effect, that we are a people who won’t put up with torture. So, it’s not true that where everyone is already refraining from behaving in a certain way, and this is common knowledge, we don’t “need” a principle forbidding individuals from behaving in that way. Such a principle may be vitally important because of what it expresses.

While there is doubtless more to be said here, there is some reason to be sceptical that deliberative contractualism is vulnerable to the particular version of the conditional fallacy that Suikkanen has in mind.[[7]](#footnote-7) In any case, there is a more straightforward way in which the conditional fallacy might appear to arise. Suppose that deliberative contractors are trying to find principles that are meant to govern the way in which certain kinds of conflicts are to be resolved. Plausibly, the principles that would make most sense for deliberative contractors to employ for the resolution of such conflicts would be principles that require resolving disagreements in much the same way that disagreements get resolved within the deliberative contractual situation: that is, by engaging in prolonged, inclusive and perfectly deliberatively rational deliberation. Given that these are the principles that it would make most sense for deliberative contractors to employ, it is natural to suppose that deliberative contractors would agree to live by such principles. If this is so, then deliberative contractualism appears to entail that we are morally required to resolve disagreements by engaging in prolonged, inclusive and perfectly deliberatively rational deliberation.

But now we have a problem. First, it is clearly not true that resolving disagreements by engaging in prolonged, inclusive and perfectly deliberatively rational deliberation is morally required of *us as we actually are*. Insisting otherwise would appear in many cases to violate the principle that “ought” implies “can.” Many of the disagreements we actually confront are subject to strict time pressures that make *prolonged* deliberation of any kind impossible. Geographical and logistical (as well as temporal) considerations will often place limits on the feasible *inclusiveness* of real-world deliberation. And many of us are sufficiently deliberatively irrational that perfectly deliberatively rational deliberation is beyond our capacities. So, at least in many cases, we would seem to be *unable* to resolve disagreements by engaging in prolonged, inclusive and perfectly deliberatively rational deliberation. But the principle that “ought” implies “can” holds, very roughly, that we ought (or are required) to do something only if we are able to do it. It follows that it is not the case that we ought (or are required) to do so.

Second, nor is it even true that *trying* to resolve disagreements by engaging in prolonged, inclusive and perfectly deliberatively rational deliberation is morally required of us as we actually are. Doing so would often have bad consequences. If we try to resolve a disagreement by engaging in prolonged and inclusive deliberation when there is insufficient time to do so, we will almost certainly not succeed in resolving the disagreement at all. If we try to approximate satisfaction of the norms of deliberative rationality, we are very likely to sacrifice the force of the better argument to the force of the loudest voice. And so on. In many cases, the best way for us to try to resolve disagreements will look *nothing like* the way that disagreements are supposed to get settled within the deliberative contractual situation.

Here, then, we have a concrete illustration of how the conditional fallacy appears to arise in the case of a particular contractualist theory, deliberative contractualism. This is not the only way the conditional fallacy potentially arises in the case of deliberative contractualism. And it will potentially arise somewhat differently in the case of different contractualist theories. But it will do for our purposes. Let us now consider a particular way that the contractualist might respond.

III. The advice model of contractualism

The response I want to explore involves appropriating on behalf of the contractualist the strategy that Michael Smith (1994; 1995) has famously deployed on behalf of the ideal observer theory. Very roughly, the ideal observer theory (IOT) holds that:

(IOT) A ought to X iff A would desire to X if A were fully informed and fully rational (Brandt 1979).

(IOT) might seem to be obviously guilty of committing the conditional fallacy (Shope 1978, pp. 411-13). To borrow a much-discussed example from Gary Watson (1975, pp. 107-9; see also Smith 1994, p. 148; Smith 1995), suppose that a particular individual (let’s call her Alma) loves playing squash but is a terribly bad sport, such that she cannot shake hands with a victorious opponent without punching her opponent in the face. It seems obvious that under these circumstances Alma ought not to shake her opponent’s hand. Rather, she ought to go and cool off. But if Alma were fully informed and fully rational, then presumably she wouldn’t be a bad sport and would desire to go and shake her opponents hand after the game. (IOT) commits the conditional fallacy because it gives the wrong answer about what Alma ought to do in her actual, non-ideal circumstances.

Smith’s response is to distinguish between what he calls the “example” and “advice” models of (IOT) and to suggest that only the former succumbs to the conditional fallacy (Smith 1995). The example model (IOTex) holds that

(IOTex) A ought to X iff if A were to be a fully informed and fully rational counterpart of herself, A\*, then A\* would desire that A\* Xs.

The advice model, by contrast, holds that

(IOTadv) A ought to X iff if A were to be a fully informed and fully rational counterpart of herself, A\*, then A\* would desire that (actual, non-ideal) A Xs.

The difference between the two models can be put simply as follows. Whereas the example model holds that the normative status of A’s actions is determined by the desires that her ideal, counterfactual counterpart A\* would have for *herself* (i.e. for A\*), the advice model holds that the normative status of A’s actions is determined by the desires that A\* would have for *actual, non-ideal* A.

How does this help the proponent of (IOT) to respond to the conditional fallacy? As we noted above, a fully informed and fully rational counterpart of Alma, Alma\*, presumably wouldn’t be a bad sport and hence would desire that she herself, Alma\*, go and shake her victorious opponents hand. This is bad news for the example model, since it entails that Alma ought to go and shake her opponent’s hand, even though this means that it is very likely that she will smack her opponent in the face. But now consider the advice model. It presumably *isn’t* true that Alma\* would desire that (actual, non-ideal) Alma go and shake her opponent’s hand. Alma\*’s desires concerning Alma’s actions would take account of Alma’s particular characteristics, including the fact that Alma is a bad sport and that she is very likely to punch her opponent in the face if she shakes his hand. So Alma\* would presumably not desire that Alma go and shake her opponents hands. Thus, the advice model does not have the objectionable consequence that Alma ought to go and shake her opponent’s hand (and can vindicate the sensible thought that she ought instead to go and cool off).

Might we appropriate Smith’s strategy on behalf of the contractualist? The idea would be to show that a) there are two different models of contractualism, an example model and an advice model and b) the conditional fallacy arises in the case of the example model but not in the case of the advice model. It doesn’t arise in the case of the advice model because the contractors would be appropriately sensitive to facts about us as we actually are; and the code by which they would agree to live would be one that was designed to take account of our inferior characteristics and circumstances. They would agree to a special real world-sensitive code *for us*—a code that would be quite different from the sort of code to which they would agree *for themselves*.

But there is a problem. The problem is that, for many versions of contractualism (including deliberative contractualism), the advice model appears to guilty of straightforward “conceptual incoherence” (Southwood 2010, p. 136, n. 24). This is because of a crucial difference between contractualism and the ideal observer theory. The ideal observer theory makes the normative status of our actions dependent on the *desires* of our ideal, counterfactual counterparts. Contractualism, by contrast, makes the normative status of our actions dependent on the *agreement* of our ideal, counterfactual counterparts. And whereas, conceptually, there is no problem with the idea that our counterfactual, ideal counterparts *desire* that we (as we actually are) perform (or refrain from performing) certain actions, it is conceptually incoherent to suppose that our counterfactual, ideal counterparts *agree* that we (as we actually are) live by particular principles. Agreements are special kinds of collective *decisions* (Southwood 2010, p. 3). And a counterpart of us cannot decide thatwe live by a principle. Only we can do that. At most, counterparts of us can decide to *encourage* us to live by a principle, to *punish* us if we deviate from the principle, perhaps even to *legislate* the principle, and so on.[[8]](#footnote-8) So it is not the case that there are two coherent models of contractualism, an example model and an advice model.

Fortunately, however, there is a solution to the problem. The alleged incoherence arises if we assume that contractors are making *unconditional decisions* about how we are to live. But there is an alternative. This is to think of the contractors as making special kinds of *conditional* decisions about the principles by which to live. We often make conditional decisions: to accept a particular job if we are offered it; to cancel tennis if it rains; to end a friendship if certain allegations made by one’s friend’s ex-wife about his past conduct are true. We might also make conditional decisions where the circumstances on which the decisions are conditional are certain kinds of considerations about *ourselves*—including considerations that will never obtain. For example, we might decide to send our mother in law a bunch of flowers for her birthday if we forget to call her, or to take some ibuprofen if we drink too much on Saturday night, or to stop being a philosopher if we suffer a brain injury.

Here, then, is how I suggest we understand the distinction between an example model and an advice model of contractualism. Both models hold that morality (or some part of it) somehow depends on some special conditional counterfactual agreement.[[9]](#footnote-9) The example model holds that morality depends on a (counterfactual) agreement to live by certain principles insofar as the contractors continue to possess *their* characteristics and occupy *their* circumstances. The advice model holds that morality depends on a (counterfactual) agreement to live by certain principles insofar as the contractors possess *our* characteristics and occupy *our* circumstances.[[10]](#footnote-10) A model of contractualism along these lines is structurally analogous to the advice model of the ideal observer theory.[[11]](#footnote-11) And it is perfectly coherent. Can it solve the conditional fallacy?

Take deliberative contractualism. Recall that deliberative contractualism holds that

(DC) It is morally impermissible/obligatory for A to X iff if we were all perfectly deliberatively rational and charged with the task of agreeing upon a common code by which to live, then we would agree to live by a common code, C, that includes a principle P that forbids/requires S to X.

The example model (DCex) holds that

(DCex) It is morally impermissible/obligatory for A to X iff if we were all counterparts of ourselves who were perfectly deliberatively rational and charged with the task of agreeing upon a common code by which to live, (A\*, B\*, …, N\*), then we (A\*, B\*, … , N\*) would agree, *insofar as our characteristics and circumstances continue to be relevantly A\*-like, B\*-like … and N\*-like*, to live by a common code, C, that includes a principle P that forbids/requires A to X.

The advice model (DCadv) holds that

(DCadv) It is morally impermissible/obligatory for A to X iff if we were all counterparts of ourselves who were perfectly deliberatively rational and charged with the task of agreeing upon a common code by which to live, (A\*, B\*, …, N\*), then we (A\*, B\*, … , N\*) would agree, *insofar as our characteristics and circumstances are relevantly A-like, B-like … , and N-like,* to live by a common code, C, that includes a principle P that forbids/requires A to X.

It seems plausible to suppose that our contractor counterparts, A\*, B\*, etc would indeed agree, insofar as their characteristics and circumstances are relevantly A\*-like, B\*-like, etc, to live by principles that would require resolving disagreements in a perfectly deliberatively rational. This is bad news for the example model since it entails that it is obligatory for us as we actually are to resolve disagreements in a perfectly deliberatively rational way, whereas this is not obligatory—and indeed it will often be impermissible to try. But now consider the advice model. It presumably isn’t true that our contractor counterparts, A\*, B\*, etc would agree, insofar as their characteristics and circumstances are relevantly A-like, B-like, etc, to live by principles that would require resolving disagreements in a perfectly deliberatively rational way. The principles that A\* and B\* would agree to live by insofar as they are relevantly A-like and B-like would take account of A’s and B’s characteristics and circumstances—including the fact that they are imperfectly deliberatively rational. So the advice model does not have the objectionable consequence that we as we actually are are obligated to resolve conflicts in a perfectly deliberatively rational way (and can vindicate the sensible though that we are obligated to resolve disagreements in completely different ways).

IV. The concessional fallacy

I now want to explore in more detail a further serious worry concerning the advice model of contractualism. This is that it succeeds in solving the conditional fallacy at the cost of being vulnerable to what I shall call the “concessional fallacy.” As we have seen, a counterfactual theory commits the conditional fallacy when it gives wrong verdicts in the actual world due to the fact that the relevant counterfactual circumstances differ in certain problematic respects from its circumstances of application (in the actual world). A theory commits the *concessional* fallacy when it gives wrong verdicts in the actual world due to the fact that the relevant counterfactual circumstances *resemble* in problematic respects, and thereby make problematic *concessions* to, its circumstances of application (in the actual world) (cf. Estlund 2014).

Again, it will be helpful to have a concrete illustration of how the concessional fallacy might arise. I shall focus on a version of the concessional fallacy that appears to arise due to the fact that many of us don’t behave as we morally ought. Take obligations of assistance. I am going to assume that there is some n such that a) those of us who are sufficiently affluent are morally obliged to donate n% of our income to assist those in dire need and b) virtually none of us actually does so. The advice version of contractualism holds that the moral status of our actions depends on the principles that contractors would agree to live by insofar as they occupy our circumstances. Our circumstances are ones in which virtually none of us donates n% of our income to those in dire need. Moreover, we can suppose that this is common knowledge among our contractor counterparts. But it might seem highly doubtful that our contractor counterparts would agree, insofar as they occupy certain circumstances, to live by principles that would require actions that they know they are virtually certain not to perform in those circumstances. So, it might seem highly doubtful that they would agree, insofar as they occupy our circumstances, to agree to live by principles that would require the affluent among them to donate n% of their income to those in dire need. Rather, we might reasonably expect them to agree, insofar as they occupy our circumstances, to live by principles that would require them to donate some less significant portion of their wealth, n-m%, to those in dire need.

But now we have a problem. The advice model of contractualism does not deliver the verdict that we are morally obliged to give n%. Rather, it delivers the verdict that we are morally obliged to give n-m%. Yet, by hypothesis, we are morally obliged to donate n%. Merely donating n-m%, while better than nothing, is nonetheless morally impermissible. So the advice model fails to deliver the correct result. The problem is that the advice model makes the moral status of our actions dependent on a specification of the contractual situation in which the contractors are tailoring principles to our actual circumstances. Our actual circumstances sometimes involve arbitrary and morally irrelevant elements. In tailoring principles to our circumstances, the contractors are tailoring principles to arbitrary and morally irrelevant considerations. The advice model appears to commit the concessional fallacy since it appears to make undue concessions to us as we actually are.[[12]](#footnote-12)

Notice that the *example* model of contractualism does not appear to be vulnerable to the concessional fallacy. That’s because it makes the moral status of our actions dependent on the principles that contractors would agree to live by insofar as they continue to occupy *their* superior circumstances. Their superior circumstances needn’t be ones where they are known to fail to give n% of their income to those in dire need. So there is no reason to suppose that the example model would make undue concessions to us as we actually are.

Nor does the advice model of the *ideal observer theory* appear vulnerable to the concessional fallacy—or at least to this version of it. Suppose that in her actual circumstances Alma won’t give n% of her income to those in dire need. While her fully informed and rational counterpart, Alma\*, may not *agree*, insofar as she occupies Alma’s actual circumstances, to live by principles that would require her to do what she won’t do, she might nonetheless *desire* that actual Alma do what she won’t do, including donating n% of her income to those in dire need. The latter, unlike the former, does not appear to involve manifesting any irrationality. At least this version of the concessional fallacy seems to be a special problem for the advice model of *contractualism* in particular.

What should we make of this objection? It might seem that to reject it, the contractualist must reject the key claim that our contractor counterparts would not agree, insofar as they occupy certain circumstances, to live by principles that would require actions that they know they won't (or are virtually certain not to) perform in those circumstances. Let us look at this claim more closely. It derives its plausibility, I take it, from the thought that there seems to be something fishy about making a decision to perform an act X in circumstances where we know we won’t (or are virtually certain not to) X. Since agreeing to live by a principle that requires us to X entails deciding to X (or so I am assuming), it follows that there is something equally fishy about agreeing to live by a principle that would require us to X in circumstances where we know we won’t (or are virtually certain not to) X.

What exactly is the nature of the “fishiness?” I can think of three possibilities. First, one might think that it is literally *impossible* to make a decision to perform an act X in circumstances C where we know we won’t (or are virtually certain not to) X in C (Harman 1997; Velleman 2000; cf. Wallace 2001). Deciding to X entails that in our mind the question of whether or not we will X remains “open.” But if we know we won’t (or are virtually certain not to) X, then the question is closed. We are not in a position even to pose the question of whether or not to X, to which the decision to X is an answer (see Hieronymi 2009; Southwood 2016a; forthcoming a; forthcoming b). Second, one might think instead that it is *irrational* to decide to do something insofar as we are in circumstances where we know we won’t (or are virtually certain not to) do it. In particular, it is contrary to the following structural requirement of rationality:

(R) We are rationally required not to decide to X in circumstances C if we believe that we won’t (or are virtually certain not to) X in C.

I would love to play cricket for the Australian cricket team. However, alas, I now believe that I am virtually certain not to play cricket for the Australian cricket team. Suppose, then, that I make a decision to play cricket for the Australian cricket team. Under these circumstances, it may seem clear that I would be irrational. Third, one might think that it is somehow *normatively inappropriate* to decide to perform an act X in circumstances C where we know we won’t (or are virtually certain not to) X in C. For example, it may seem pointless or a waste of time to do so. Alternatively, it might seem to be insincere or irresponsible, if not to make the decision, then at least to communicate it to others.

None of these three claims is beyond question. But let’s simply concede that making a decision to perform an act X in circumstances where we know we won’t (or are virtually certain not to) X—and, hence, agreeing to live by a principle that would require us to X in those circumstances—is indeed fishy in one (or more than one) of these ways. And let’s concede that this is enough to vindicate the key claim that our contractor counterparts would not agree, insofar as they occupy certain circumstances, to live by principles that would require actions that they know they won't (or are virtually certain not to) perform in those circumstances. It might seem to follow straightforwardly that our contractor counterparts would not agree, insofar as their characteristics and circumstances are relevantly like ours, to live by principles that would require us to do things (such as donate n% of our income to those in dire need) that we won’t (or are virtually certain not to) do.

But it doesn’t follow at all. For there is a crucial difference between a) the claim that given our actual characteristics and circumstances we won’t (or are virtually certain not to) perform some act X and b) the claim that given our actual characteristics and circumstances we wouldn’t (or would be virtually certain not to) perform X *even if were to agree to live by principles that would require us to X*. In order to show that our contractor counterparts would not agree, insofar as their characteristics and circumstances are relevantly like ours, to live by principles that would require us to perform X, then, it is not enough to show that our contractor counterparts would know that (a). Rather, the critic must show that they would know that (b).

There are certainly acts for which (b) is plausible. Suppose that we could prevent a genocide by, say, bowling a cricket ball at 160 kilometres an hour, or playing Ravel’s Gaspard’s de la nuit (flawlessly and at the correct tempo), or calculating the square root of 13257076354384052 within three seconds. Alas, most of us are unable to do any of these things. Plausibly, this entails that we are robustly disposed, insofar as we form an intention or make a decision to do any of these things, to fail. Presumably our contractor counterparts would be in a position to know this. In consequence, it seems reasonable to surmise that they would not agree to live by principles that would require us to bowl a cricket ball at 160 kilometres an hour, or to play Ravel’s Gaspard’s de la nuit, or to calculate the square root of 13257076354384052 within three seconds.

However, I take it that acts that we are unable to perform are not especially promising candidates for establishing that contractualism is vulnerable to the concessional fallacy. The problem is that insisting that we are morally required to perform such acts is at odds with the idea that “ought” implies “can.” To be sure, “ought” implies “can” is not uncontroversial; indeed, one important source of resistance to it is precisely that it appears to license unduly concessive normative inferences (Stocker 1971, pp. 314-5; Sinnott-Armstrong 1984, pp. 252-4; Young 1975, pp. 13-14). Nonetheless, I’m going to assume that relying upon cases that require us to give up on the idea that “ought” implies “can” is simply too high a cost.[[13]](#footnote-13)

Suppose instead, then, that the act in question, rather than being an act that we are unable to perform, is an act that we simply won’t (or at least are virtually certain not to) perform. So far as I can tell there are just two ways this might happen. First, suppose that we won’t (or are virtually certain not to) perform the act because we won’t (or are virtually certain not to) *decide* to perform it. For example, I am extremely unlikely to abandon philosophy to become a labourer because I’m extremely unlikely to decide to abandon philosophy to become a labourer. This does not mean that I would be unlikely to abandon philosophy to become a labourer if I were to decide to do so. On the contrary, the only reason to think that I won’t abandon philosophy to become a labourer is that I am extremely unlikely to decide to do so. If I were to decide to abandon philosophy to become a labourer, there is every reason to think that that is exactly what I would do. A fortiori, the chances are that I would do so if I were to agree to live by principles that would require me to abandon philosophy to become a labourer.

Second, suppose that, even if we decide to perform the act, we won’t (or are virtually certain not to) perform the act because we won’t (or are virtually certain not to) *try* (or keep trying) to perform the act. For example, I am extremely unlikely to read the complete works of Alain Robbe-Grillet because I’m extremely unlikely to make a genuine attempt to do so. Even if I were to reach a decision to do so, I find the *nouveau roman* to be such an insufferably dull genre that I would be extremely unlikely to sustain the effort required to wade through so much repetitive prose. Does this mean that I would be extremely unlikely to read the complete works of Alain Robbe-Grillet if I were to agree to live by principles that would require me to do so? Certainly not. For one, agreements are *collective* decisions. As such, they involve others. Even if I would fail to read the complete works of Alain Robbe-Grillet if I were to decide (on my own) to do so, it does not follow that I would fail to do so if *you and I* were to decide to do so. For another, agreeing to live by a principle that requires one to X involves more than deciding to X. It also plausibly involves internalising the principle. While it is a matter of controversy precisely what internalising a principle involves, plausibly part of what it involves is precisely a disposition to try to comply with it. So, again, even if I would fail to read the complete works of Alain Robbe-Grillet if I (or even you and) were to decide to do, it does not follow that I would fail to do so if I were to internalise a principle that would require me to do so.

We are now in a position to respond to the charge that the advice model of contractualism commits the concessional fallacy. To vindicate the charge, the critic must show that there is some act (e.g. donating n% of our income to those in dire need) such that 1) given our actual characteristics and circumstances, we wouldn’t (or would be virtually certain not to) perform the act if we were to agree to live by principles that would require us to do so; and 2) we are morally required to perform the act. The problem is that it seems that the only reason to think that an act has feature (1) is that we are unable to perform the act, which means that it lacks feature (2) —on the assumption that “ought” implies “can.” If instead we are able to perform the act, then the fact that we won’t (or are virtually certain not to) perform the act is not a reason to think that we wouldn’t (or would be virtually certain not to) perform the act if we were to agree to live by principles that would require us to do so. Thus, the fact that we won’t (or are virtually certain not to) perform an act such as donating n% of our income to those in dire need is not a reason to think that the advice model of contractualism will have any difficulty delivering the correct verdict that we are morally required to do so.

V. Obligations to do the next best thing

Let me briefly mention a residual concern. The concern is that our response to the concessional fallacy charge, if successful, makes the advice model incapable of explaining moral obligations to do the next best thing. Consider two individuals, A and B, neither of whom gives n% of her income to those in dire need. But suppose that whereas A gives nothing at all, B gives n-m%. There seems to be an important moral difference between A and B. The difference isn’t just that B comes closer to fulfilling her moral obligation to donate n%. Rather, it’s that, while, to be sure, both A and B act wrongly inasmuch as they fail to comply with the moral obligation to give n%, B, unlike A, nonetheless seems to get *something* right. The problem is that, to the extent that our response to the concessional fallacy charge is successful, it isn’t obvious how the advice model can explain this. If our contractor counterparts would agree to principles that would require us to give n%, it seems clear that they would not agree to give n-m%, since giving n-m% entails not giving n%.

Here is another example. We are morally obligated not to shout at our spouses (absent special circumstances). Alas, most of us do sometimes shout as out spouses. When we do so we behave wrongly. But now consider two individuals, C and D, both of whom shout at their spouses. But suppose that whereas C continues to be beastly, D apologises afterwards and makes an effort to be especially kind and caring. Both C and D do something wrong, since they both shout at their spouses; and shouting at your spouse is wrong. But they are not morally on par. D, unlike, C seems to get something right by apologising and being especially kind and caring afterwards. By contrast, C seems to behave wrongly twice over: first, in shouting at her spouse; second, in not apologising and continuing to be beastly. Again, to the extent that the advice model can explain the obligation not to shout at our spouses (in spite of the fact that most of us violate the obligation) it’s not clear how it can explain the moral difference between C and D.

In fact, I believe that there is a relatively simple response to this concern. The key is to note that the principles by which our contractor counterparts would agree to live, insofar as they possess characteristics and occupy circumstances relevantly like ours, fall into two categories. The first simply tell us what we are obliged to do. Call these *unconditional principles*. Here are two examples of unconditional principles:

(P1) The sufficiently affluent among us are obliged to give n% of our income to those in dire need.

(P2) We are obligated not to shout at our spouse.

But there are also principles that tell us what we are obliged to do if we don’t behave in accordance with unconditional principles. Call these *conditional principles*. Here are two examples of conditional principles:

(P3) The sufficiently affluent among us are obliged to give n-m% of our income to those in dire need if we don’t give n% of our income to those in dire need.

(P4) We are obligated to apologise and be especially nice afterwards if we shout at our spouse.

I suggest that our contractor counterparts would agree to live by principles of both kinds. This is enough to explain the moral difference between A and B and the moral difference between C and D.

In saying this, it is important to avoid a couple of potential errors. The first potential error involves assuming that the advice model holds that *all* the principles by which our contractor counterparts would agree to live are conditional principles: principles that require certain conduct insofar as we possess certain characteristics and occupy certain circumstances. That’s simply not right. There is a crucial difference between a conditional agreement to live by a principle and an agreement to live by a conditional principle. The advice model holds the moral status of our conduct depends on principles by which our contractor counterparts would conditionally agree to live. Some of these principles are unconditional. Others are conditional: that is, they tell us what we are morally obliged to do if we fail to comply with unconditional principles.

The second potential error involves the idea that adducing conditional (as well as unconditional) principles could provide an alternative, fully adequate (and rather simpler) solution to the conditional fallacy – and, hence, remove the need for the advice model. The idea would be that we interpret contractualism as holding that the moral status of our conduct depends on unconditional agreements to live by conditional (as well as unconditional) principles. For example, an interpretation of deliberative contractualism along these lines would seem to be capable of vindicating the claim that we are morally required to solve disagreements in ways that make sense given our inferior characteristics and circumstances. That’s because our deliberatively rational contractor counterparts would presumably agree to live by the following conditional principle: we are obliged to solve disagreements in some non-deliberatively rational way W given that we won’t solve disagreements in a fully deliberative way.

Nonetheless, a view of this kind is obviously not capable of solving both the conditional and the concessional fallacy. Suppose, first, that we interpret contractualism as holding that the moral status of our conduct depends on unconditional agreements to live by conditional *and unconditional* principles. In that case, while it may entail that i) we are obliged to solve disagreements in some non-deliberatively rational way W given that we won’t solve disagreements in a fully deliberative way, it will presumably also entail that ii) we are morally required to solve disagreements in a fully deliberative way. So it commits the conditional fallacy. Suppose instead that we interpret contractualism as holding that the moral status of our conduct depends on unconditional agreements to live by conditional principles alone: principles that require us to perform certain acts given that we will and won’t perform certain acts. In that case, it will be guilty of committing the concessional fallacy. For it will be incapable of vindicating the claim that we are morally required to donate n% of our income to those in dire need. In sum, the advice model is not dispensable. We need both conditional agreements and conditional (as well as unconditional) principles.

VI. Conclusion: the ideal/non-ideal theory issue

I have proposed a model of contractualism that is structurally analogous to the advice model of the ideal observer theory. I have suggested that an advice model of contractualism is perfectly coherent once we interpret it in the right way. And I have argued that it plausibly avoids both the conditional and concessional fallacies. I want to conclude by saying something about what this might mean for the so-called “ideal/non-theory” issue.

The ideal/non-ideal theory issue is not well defined. But I take it that at least one animating concern of participants in the debate is the following: What is the (relative) importance of, and relation between, two canonical modes of normative theorising: one that is supposed to be utopian, based on unrealistic assumption and avowedly insensitive to certain features of the actual world; the other that is supposed to be non-utopian, based on realistic assumption, and avowedly sensitive to certain features of the actual world (see Valentini 2012).[[14]](#footnote-14)

The current paper connects with this concern in at least three ways. First, it offers a way of precisifying certain common (but rather vague) complaints that are made by participants within the debate regarding their opponents. On the one side, it is often claimed that there is something objectionable about the view that utopian modes of normative theorising are explanatorily prior to, and helpful (perhaps even indispensable) for, engaging in non-utopian modes of normative theorising. One thing this complaint might be is that certain versions of this view (or perhaps even the view as such) are guilty of committing the conditional fallacy. On the other side, it is often claimed that there is something objectionable about engaging in non-utopian modes of normative theorising in isolation from utopian modes of normative theorising. One thing this complaint might amount to is that certain versions of this view (or perhaps even the view as such) are guilty of committing the concessional fallacy.

Second, it offers *answers* to these common complaints, thus construed. For one, the advice model of contractualism is an example of a utopian normative theory that is based on unapologetically unrealistic assumptions. Yet, I have also argued that it is a mistake to suppose that it entails principles that don’t make sense for us as we actually are. This shows that the complaint that utopian theorising will produce principles that don’t make sense for us as we actually are is misguided. For another, the advice model of contractualism is an example of normative theory that in certain respects is based on, and tailored to, our actual characteristics and circumstances. Yet, I have argued that it is a mistake to suppose that it makes undue concessions to us as we actually are. This shows that the complaint that non-utopian theorising will generate undue concessions to us as we actually are is misguided.

Third, and most importantly, it shows that there is something misguided about the ideal/non-ideal theory debate—or at least something misguided about the animating concern that I mentioned above. For the debate appears to assume that there are just two distinct canonical modes of normative theorising—a utopian mode and a non-utopian mode—and that political philosophy will have to involve one or both of them in some combination. Against this, notice that the advice model of contractualism is not an example of either mode or a combination of them. Rather, it represents a third way. It is an instance of a mode of normative theorising that is unrealistic and utopian in the sense that the characteristics and circumstances of the contractors are drastically at odds with our actual characteristics and circumstances. Yet it is also highly realistic and non-utopian in the sense that the characteristics and circumstances of the individuals for whom the contractors are seeking to find principles are just those of us as we actually are.

References

Anderson, Elizabeth and Richard Pildes. 2000. “Expressive theories of law: a general restatement,” *University of Pennsylvania Law Review*, 148, 1503-75.

Ashford, Elizabeth. 2003. “The demandingness of Scanlon’s contractualism,” *Ethics*, 113, 273-302.

Ashford, Elizabeth and Tim Mulgan. 2012. “Contractualism,” *Stanford Encyclopedia of Philosophy*, available at https://plato.stanford.edu/entries/contractualism/

Brennan, Geoffrey, Lina Eriksson, Robert E. Goodin and Nicholas Southwood. 2013. *Explaining Norms*. Oxford: Oxford University Press.

Darwall, Stephen. 2003. Introduction. Pp. 1–8 in Stephen Darwall (ed.), *Contractarianism/Contractualism*. Oxford: Blackwell Publishing.

Estlund, David. 2014. “Utopophobia,” *Philosophy & Public Affairs*, 42, 113-34.

Forst, Rainer. 2014. *The Right to Justification*. New York: Columbia University Press.

Habermas, Jürgen. 1984–1987. *The Theory of Communicative Action*, trans. Thomas McCarthy. Boston, MA: Beacon Press.

Hampton, Jean. 1986. *Hobbes and the Social Contract Tradition*. New York; Cambridge, UK: Cambridge University Press.

Harman, Gilbert. 1975. Moral relativism defended. *Philosophical Review*, 84, 3–22.

Harman, Gilbert. 1997. “Practical Reasoning.” In *The Philosophy of Action*, ed. Alfred Mele, Oxford: Oxford University Press, 149-77.

Gauthier, David. 1986. *Morals By Agreement*. Oxford: Clarendon Press.

Hieronymi, Pamela. 2009. ‘The Will as Reason’, *Philosophical Perspectives*, 23, 201-220.

Hobbes, Thomas. 1994. *Leviathan*. Indianapolis, IN: Hackett.

Hooker, Bard. 2000. *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford: Clarendon Press.

Kavka, Gregory. 1986. *Hobbesian Moral and Political Theory*. Princeton, NJ: Princeton University Press.

McGinn, Colin. 1999. Reasons and unreasons. *The New Republic*, 24 May, 34–8.

Moore, G. E. 1912. *Ethics*. London: Williams and Norgate.

Muldoon, Ryan. 2016. *Social Contract Theory for a Diverse World: Beyond Toleration*. Abingdon: Routledge.

Scanlon, T. M. 1998. *What We Owe To Each Other*. Cambridge, MA: Belknap Press of Harvard University Press.

Schroeder, Mark. 2011. “Ought, Agents, and Actions,” *The Philosophical Review*, 120, 1-41.

Shope, Robert. 1978. “The conditional fallacy in contemporary philosophy,” *Journal of Philosophy*, 75, 397-413.

Sinnott-Armstrong, Walter. 1984. “‘Ought’” conversationally implies ‘can,’” *Philosophical Review*, 93, 249-61.

Smith, Michael. 1994. *The Moral Problem*. Oxford; Cambridge, MA: Blackwell.

Smith, Michael. 1995. “Internal Reasons,” *Philosophy and Phenomenological Research*, 55, 110-31.

Southwood, Nicholas. 2009. “Moral Contractualism, *Philosophy Compass*, 4, 926-937.

Southwood, Nicholas. 2010. *Contractualism and the Foundations of Morality*. Oxford: Oxford University Press.

Southwood, Nicholas. 2016a. ‘“The Thing to Do” Implies “Can”’, *Noûs*, 50, 61-72.

Southwood, Nicholas. 2016b. “Does ‘Ought’ Imply ‘Feasible’?” *Philosophy & Public Affairs*, 44, 7-45.

Southwood, Nicholas. Forthcoming a. The question of practical reason, *Reasoning: Essays on Theoretical and Practical Thinking*, ed. M. Balcerak-Jackson and B. Balcerak-Jackson. Oxford: Oxford University Press.

Southwood, Nicholas. Forthcoming b. “Constructivism and the Normativity of Practical Reason,” *The Many Moral Rationalisms*, ed. K. Jones and F. Schroeter. Oxford: Oxford University Press.

Stocker, Michael. 1971. “‘Ought’ and ‘can,’” *Australasian Journal of Philosophy*, 49, 303-16.

Suikkanen, Jussi. 2014. “Contractualism and the conditional fallacy,” *Oxford Studies in Normative Ethics*, 4, 113-37.

Thompson, Judith Jarvis. 1990. *The Realm of Rights*. Cambridge, MA: Harvard University Press.

Valentini, Laura. 2012. “Ideal versus non-ideal theory: a conceptual map,” *Philosophy Compass*, 7, 654-64.

Velleman, David. 2000. *The Possibility of Practical Reason*. Ann Arbor: University of Michigan Press.

Wallace, R. Jay. 2001. ‘Normativity, Commitment, and Instrumental Reason’, *Philosophers’ Imprint*, 1/3 (December).

Watson, Gary. 1982. “Free Agency,” *Free Will*, ed. Gary Watson. Oxford: Oxford University Press.

Wolf, Susan. 1990. *Freedom Within Reason*. New York: Oxford University Press.

Young, Robert. 1975. *Freedom, Responsibility and God*. New York: Harper and Row.

1. For helpful comments, criticism, discussion, and encouragement I would like to thank Dirk Baltzly, Christian Barry, Bob Goodin, Jussi Suikkanen, audiences at the Future of Contractualism Conference at the University of Rennes, the University of Melbourne, Monash University, and the Beyond Contractarianism conference at the Inter-University Centre in Dubrovnik organised by Ohio State University, and an anonymous referee for this journal. I am especially grateful to Jussi Suikkanen for raising the conditional fallacy problem for my preferred “deliberative” model of contractualism. Research for the article was supported by an Australian Research Council-funded Future Fellowship (FT160100409). [↑](#footnote-ref-1)
2. Southwood 2009; 2010. Notice that this characterization of contractualism in broader than certain other characterizations of contractualism, such as Darwall’s (2003) and Ashford and Mulgan’s (2008). [↑](#footnote-ref-2)
3. Thomas Hobbes (1994), for example, often speaks of the agreement to establish an omnipotent sovereign, whose dictates are constitutive of justice, as an actual historical event (though see Hobbes 1994, p. 109). Similarly, Jurgen Habermas (1984-1987) sometimes appears to suggest that morality and justice are determined by actual deliberation. Gilbert Harman (1975, p. 3) writes that “morality arises when a group of people reach an implicit agreement or come to a tacit understanding about their relations with one another.” Most recently, Ryan Muldoon (2016, p. 120) has also defended the idea that justice is the product of a series of actual, ongoing bargains for the sake of mutual advantage. [↑](#footnote-ref-3)
4. A trivial moral truth is a true claim of the form “X is permissible” that is true simply because there are no true claims of the form “X is impermissible.” [↑](#footnote-ref-4)
5. Surprisingly, with one notable exception, contractualists have not engaged much with the conditional fallacy directly. The exception is Jussi Suikkanen (2014). In a recent paper, Suikkanen explicitly considers the question and argues that two counterfactual versions of contractualism, the “deliberative contractualism” that I have myself defended (Southwood 2010) and T. M. Scanlon’s “Kantian contractualism” (Scanlon 1998), are indeed guilty of committing the conditional fallacy. [↑](#footnote-ref-5)
6. It might be objected that the probability is zero as long as we satisfy the description of the situation. But the point is that there is non-zero probability that we will continue to satisfy that description. I am grateful to Jussi Suikkanen for this nice way of putting the point. [↑](#footnote-ref-6)
7. Suikkanen (2014, pp. 124-26) offers some interesting lines of response, which I won’t try to address here. It is worth emphasising, however, that if the “advice model” of contractualism succeeds in avoiding the alternative version of the conditional fallacy considered below, then it also will succeed in solving Suikkanen’s version. [↑](#footnote-ref-7)
8. This might be resisted on the grounds that others can make decisions about how we are to act when they are *authorised* by us to make decisions on our behalf. But this doesn’t help the deliberative contractualist. That’s because none of us has authorised our ideally deliberatively rational counterparts to make decisions about what common code we are to live by. In that case, it remains impossible for our ideally deliberatively rational counterparts to decide that we are to live by certain principles. Notice, moreover, that if the deliberative contractualist is forced to go down this path, then she will confront anew some of the troublesome aspects of actualist contractualism that recourse to counterfactualist contractualism was supposed to avoid. In order to be normatively significant, an act of authorisation must presumably satisfy stringent constraints. But it is unrealistic to suppose that these constraints will be universally satisfied in the actual world. [↑](#footnote-ref-8)
9. What about a model of contractualism that holds that morality depends on some *unconditional* counterfactual agreement: that is, an agreement to live by certain principles *irrespective of the contractors’ characteristics and circumstances*? This is clearly an option in conceptual space. But it is a relatively uninteresting one. Given, as we have seen, whether a particular principle makes sense is heavily dependent on the characteristics and circumstances of the agent whose conduct it is supposed to regulate, it would seem likely that conscientious contractors would be sensitive to these considerations. [↑](#footnote-ref-9)
10. This is also important to foreclose a possible worry with interpreting contractualism in light of conditional agreements: that doing so means that we will have an objectionably high degree of freedom in specifying what the contractors would agree to. If the agreements of the hypothetical contractors are meant to be binding, why wouldn't our disputes now simply turn on which of several sets of plausible conditional agreements would be agreed to by our advice-givers? The answer to this worry is that the *relevant* agreements are those that are conditional on how we actually are. I am grateful to Jussi Suikkanen for raising this worry. [↑](#footnote-ref-10)
11. It might be objected that there is something a little misleading about describing it as an “advice model” since a conditional decision to X is quite different from advice to X. That’s right. But it is no more misleading than Smith’s description of his theory as an advice model of the ideal observer theory. Having a desire that S performs X is also quite clearly different from advice to S to X. I have opted to call it “the advice model of contractualism” simply because it is structurally analogous to the theory that Smith calls “the advice model of the ideal observer theory” —a label that, while misleading, is deeply entrenched in philosophical usage. [↑](#footnote-ref-11)
12. In personal correspondence, Jussi Suikkanen has made the interesting suggestion that perhaps “the advisors would agree to give more than n% in the non-ideal circumstances. They might think that, even if most of them will end up giving less than n, insofar as they are able to live by their decision they would want to do more given how many people there are suffering. … This of course doesn’t change the problem. It would just entail that the code would be too demanding rather than not demanding enough.” This seems to me to be quite plausible as a claim about Scanlonian contractualism in particular (see Ashford 2003) but much less plausible as a claim about other kinds of contractualism. At the very least, whereas there is no *oddity* in the idea that a principle that would require us to do what none of us will do couldn’t reasonably be rejected, there *is* a serious prima facie oddity in the idea that we would agree to live by a principle that would require us to do what none of us will do (see below). [↑](#footnote-ref-12)
13. Moreover, while there are putative counterexamples to the idea that “ought” implies “can,” no one to my mind has developed a remotely credible counterexample to the idea that the so-called “deliberative ought” implies “can.” The deliberative ought is distinctive inasmuch as claims involving it are supposed to be fit to be used in settling the practical question of what to do. It is very hard to see how ought claims of this kind could be true in cases when agents cannot perform them (Schroeder 2011; Southwood 2016b). And contractualism (of the kind that is my subject here) plausibly aspires to explain truths that entail or at least contribute to corresponding truths involving the deliberative ought. [↑](#footnote-ref-13)
14. There are many different positions we might take. But three relatively common positions (I take it) are a) that both modes of normative theorising are important but that the utopian mode is explanatorily prior to the non-utopian mode, b) that the non-utopian mode is important and wholly independent of the utopian mode, which is relatively unimportant c) that both modes are important (for different purposes) and neither is explanatorily prior to the other. [↑](#footnote-ref-14)