May 2015

Two Contemporary Metaethical Schemes Considered

Benjamin Serber
University of Wisconsin-Milwaukee

TWO CONTEMPORARY METAETHICAL SCHEMES CONSIDERED

by

Benjamin Serber

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Arts

in Philosophy

at

The University of Wisconsin Milwaukee

May 2015

ABSTRACT
TWO CONTEMPORARY METAETHICAL SCHEMES CONSIDERED

by

Benjamin Serber

The University of Wisconsin Milwaukee, May 2015
Under the Supervision of Professor Stanislaus Husi

This thesis examines two of the more modern developments in the field of metaethics, expressivism and rational choice metaethics. Metaethics deals with a number of questions surrounding what we actually do when we engage in moral thought and speak in moral language. I approach the debate through the question of the objects of moral language. As metaethics has diversified away from straightforward moral realism, a number of candidates have been proposed as the actual referents of the moral terms we use. In expressivism, the object of moral language is taken to be certain nonpropositional attitudes held by the speaker of a morally tinged sentence. In rational choice metaethics, the object of moral language is taken to be choices made by agents, evaluated through the lens of rational choice theory. I examine expressivism and rational choice metaethics, concluding that both retain serious problems. For expressivism, I attempt to defend against a serious semantic issue, the negation problem, proposing a solution based on what I call implicit planning. Ultimately, I find the negation problem cannot be evaded and that expressivism is therefore unworkable. For rational choice metaethics, I show that its creator, David Gauthier, does not believe that the view can support some of our moral intuitions, even though all of them are meant to be derivable from its starting principles. For this reason, I find rational choice metaethics also unsuccessful.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

The author would like to thank the following individuals for their assistance with this thesis and for their parts in his graduate career:
Stan Husi, advisor
Miren Boehm, almost-advisor and committee member
Robert Schwartz, committee member and pragmatist indoctrinator
Joshua Spencer, for invaluable feedback on early drafts
The faculty of the UWM philosophy department
The graduate students – commiserators, confabulators, interlocutors, friends – of the UWM philosophy department

0.0 Introduction

This thesis will consider two of the more modern developments in the field of metaethics, expressivism and rational choice metaethics. Metaethics deals with a number of questions surrounding what we are actually doing when we engage in moral thought and speak in moral language. The particular question I will use to approach the debate is the question of the objects of moral language. As metaethical debate has diversified away from straightforward moral realism, a number of candidates have been proposed as the actual referents of the moral terms we use. In expressivism, the object of moral language is taken to be certain nonpropositional attitudes held by the speaker of a morally tinged sentence. In rational choice metaethics, the object of moral language is taken to be choices made by agents, evaluated through the lens of rational choice theory. Below, I will examine first expressivism and then rational choice metaethics, concluding that both retain serious problems that should place their adoption in doubt. For expressivism, I attempt to defend against a serious semantic issue, the negation problem, proposing a solution based on what I call implicit planning. Ultimately, I find the negation problem cannot be evaded and that expressivism is therefore unworkable. For rational choice metaethics, I show that its creator, David Gauthier, does not believe that the view can support some of our moral intuitions, even though all of them are meant to be derivable from its starting principles. For this reason, I find rational choice metaethics also unsuccessful. I will now begin the discussion of expressivism in greater detail.

1.0 Implicit Planning and the Embedding/Negation Problem

One of the central questions of metaethics revolves around the nature and existence of moral facts. Some metaethicists are moral realists. They believe that the

objects of moral language are properties that actions have that make them right or wrong. This view has been subject to a number of criticisms, including G. E. Moore's Open Question argument (Sayre-McCord 2011). Other metaethicists are moral antirealists, who believe that there are no features of the world that license our normative claims. This view is often criticized as leading to some sort of subjectivism about morality (Joyce 2009). A third group of metaethicists are expressivists.[1] They argue that the objects of moral language are expressions of a nonpropositional attitude[2] held by the speaker. They also argue these attitudes can still provide normative 'oomph' and warrant the same kinds of moral talk as realist semantics that are predicated on the actual existence of moral facts. Expressivism is intended to avoid the worst objections to both moral realism and moral anti-realism. Talk of moral facts in the world is gone, addressing the largest worries with moral realism. And expressivists think they have a structure that will warrant the forcefulness of moral statements and avoid the concern that without realism our moral pronouncements are entirely subjective. If successful, this would be a very valuable metaethical project.

Many criticisms have been levelled against expressivism. One class of criticism argues that expressivism cannot adequately capture the semantics that moral realism can offer. Since one of the aims of expressivism is to continue warranting the use of realist semantics, a failure to capture realist semantics would be a major problem for the expressivist project. I aim to defend Allan Gibbard's plan-based version of expressivist

[1]More precisely, expressivism is a particular kind of metaethical quasi-realism. My focus here is only on expressivism, however.

[2] Expressivists differ on what attitude is expressed by moral statements. The views of the authors I discuss will be detailed below.

semantics as found in *Thinking How to Live* (2003) against one such semantic criticism, a version of the negation problem raised by Nicholas Unwin[3] and James Dreier,[4] which derives from the embedding problem attributed to Peter Geach and Gottlob Frege. Both problems focus on difficulties with using nonpropositional expressions of attitudes in common propositional sentences. The embedding problem raises the issue for conditional sentences, while the negation problem does the same for negation. If expressivist semantics fails to provide a sensible way of negating normative claims, this will demonstrate that expressivism is incapable of capturing significant aspects of moral language that moral realism can cover.

Overall, my endeavor proceeds as follows. Section I will cover the classic embedding problem and show how Gibbard's plan-based semantics solves the difficulty of embedding expressions into sentences with connectives. I then turn to the negation problem in Section II and explain how Gibbard's plans appear to give the expressivist a route to capture each of three key ways of negating an attitude. However, this approach will require Gibbard to also adopt the idea of a hyperplan, a meta-planning state which covers all eventualities. In Section III, I further discuss hyperplans and note a dilemma for Gibbard's view raised by Dreier. The hyperplan view prevents Gibbard from capturing certain parts of realist semantics, but abandoning hyperplans leaves him unable to capture other parts. In Section IV, I offer a view of implicit planning which addresses Dreier's concern. I take it as part of human experience that our presence in a situation entails that we are forming a plan about what to do in that situation. The plan may be

[3] In "Norms and Negation: A Problem for Gibbard's Logic," 2001.
[4] In "Negation for Expressivists," 2006.

nebulous, but it is always present. With implicit planning in hand, I argue, the expressivist will be able to differentiate between the three types of negation without running into Dreier's worry. Ultimately, however, I determine that my proposed solution does not fully avoid the negation problem, and for this and other reasons expressivism ought to be abandoned.

1.1 – The Embedding Problem

I will begin with a quick sketch of the embedding problem in its most current form.[5] The basic thrust of the problem is that expressions of attitudes just can't fit into a logical scheme the same way that propositions do. Because expressivists want to preserve everyday moral discourse (which generally sounds realist), the attitudes expressed on their view need to be able to fit into realist-sounding sentences. Accordingly, the statement "φ is wrong," when read as an expression of an attitude, needs to be able to follow the same inference rules as the statement does when read as an assertion of a proposition. The moral realist wants to say that any attempt to take all the standard rules of logic, replace some propositions with expressions of attitudes, and expect everything to continue working as before is doomed to failure. The original formulation of the embedding problem offers such an embedding of an attitudinal expression into a piece of propositional logic, the modus ponens:

R1) φ is wrong → encouraging someone to φ is wrong.
R2) φ is wrong.
RC) Encouraging someone to φ is wrong.

This is a plausibly common argument that could take place in an everyday ethical discussion, so an expressivist will definitely need a way to make the inference go

[5] I broadly follow van Roojen 1996 and Unwin 1999.

through. On a realist picture, it would do so without any trouble. In order for the inference to work, the second premise must be identical with the antecedent of the first premise; this will then satisfy the conditional and license the inference. And for the realist this is the case. R2 is the same proposition as the antecedent of R1. For the expressivist, things are more complicated. Let's take a simple expressivist view: we'll say that the statement "φ is wrong" expresses the attitude "Boo φ!"[6] So now we would have:

E1) Boo φ! → Boo encouraging someone to φ!
E2) Boo φ!
EC) Boo encouraging someone to φ!

Now for the inference to hold, E2 and the antecedent of E1 need to express the same attitude. But E1 doesn't seem to *express* an attitude at all. I can assent to or assert E1 without having either of the attitudes within it. In fact, I can assent to or assert E1 regardless of my attitude towards doing φ. If any of the parts of E1 expressed an attitude, then E1 as a whole would express an attitude as well. Since it doesn't, it can't be the case that the antecedent of E1 expresses an attitude. And since the antecedent of E1 doesn't express an attitude, it can't be identical with the attitude expressed by E2. But if that's the case, the antecedent of the conditional in E1 is not satisfied, and thus the conclusion no longer holds. Expressivists look unable to license modus ponens using expressions of attitudes, and it looks possible to generalize the problem out to other logical operators and inference rules. This conclusion puts expressivists in a dilemma. Either "φ is wrong" is an expression of a nonpropositional attitude but expressivism cannot capture critical elements of our everyday talk, or it can capture everyday talk but must give up on the idea that normative statements are expressions of attitudes. In either case, a central aim of

[6] We can symbolize this as B!φ. Conversely, "φ is right" can be "Hooray φ!" and symbolized H!φ. This notation is drawn from Blackburn.

the expressivist project will have failed. Gibbard's planning semantics, however, look to offer a solution to this problem, as indeed they are designed to do.

Gibbard offers his solution to the embedding problem by transforming the conditional of P1 into a disjunction and offering a different view of the attitudes expressed by our normative statements. For Gibbard, such a normative statement is the expression of a planning state. When someone says that an action is right, they are expressing that they have a plan to take that action, and perhaps also urging others to take that action. Turning the embedding problem into a disjunction and applying Gibbard's view of attitude expressions as planning states yields the following:

E1′) Plan to φ ∨ do not plan to encourage someone to φ.
E2′) Do not plan to φ.
EC′) Do not plan to encourage someone to φ.

While E1′ is logically equivalent to E1, Gibbard argues that expressing it as a disjunction shows that an inferential pressure remains. As a general logical principle, if I assent to a disjunction, I must accept at least one of the disjuncts. In expressivist terms, to assent to E1′, I will need to have at least one of the two attitudes it incorporates. This principle will hold even if E1′ does not express an attitude itself or contain an attitude towards φ. Accepting E1′ rules out my having (and therefore also my expressing) every combination of attitudes that does not include at least one of the two disjuncts. When viewing my attitudes as planning states in the way Gibbard does, the disjunction creates a constraint on which actions I can plan to take while remaining consistent in my overall commitments. If I accept E1′, I must be either planning to φ or not planning to encourage someone else to φ. The argument's conclusion is now licensed in a way that it was not before. Since I have committed to not rejecting both elements of the E1′ disjunct, and

with E2′ I have rejected one element, the only way to preserve my commitment to E1′ overall is to accept the conclusion.

Once the picture is reconceptualized in this way, we can see how attitudes as planning states can slot into classical semantics. Both rule out particular states of affairs. To further build up this structure and get access to counterfactual semantics, Gibbard sets up expressivist semantics as analogous to realist possible-world semantics. There is some set of possible worlds in which there are different combinations of facts and in which I hold different combinations of attitudes. In realist semantics, when I assert a proposition I am making a claim about which world(s) I might be in. I am ruling out my being in any world in which the proposition I assert is false. Gibbard expands this to note that the ruling out can apply to attitudes that I hold in addition to properties of the world. When I assent to a statement like E1′, I am ruling out being in any world in which I do not have at least one of the two disjuncted attitudes. When I accept the inference, I am thereby ruling out being in any world in which I plan to encourage someone else to φ. And since the attitudes that I hold about right and wrong are states of planning or not planning to take certain actions, the attitudes have a real impact on the sets of worlds ruled out by their acceptance. Simple booing and hooraying can fail to grab onto the world, but a plan to act carries with it a commitment to take the planned action.

This looks like a success for Gibbard. Plan-based semantics and a reorientation around disjunction and ruling out certain combinations of facts and attitudes provides a way to license realist-sounding inferences without committing to moral realism. This defends Gibbard from the semantic argument that expressivism will not be able to account for critical features of our everyday moral discourse. Additionally, the fact that

planning states have a motivational component means that I'm not just emptily expressing a feeling when I say that φ is right. When I say φ is right (which means that I plan to φ), I am ruling out every combination of facts and plans in which I do not plan to φ. This is a commitment that then motivates me to actually φ if I want to be consistent and rational. All in all, Gibbard's plan-based expressivism appears to successfully reach the goal of realist semantics without realist metaphysics.

1.2 – The Negation Problem

Unfortunately, while Gibbard's plans seem to solve the embedding problem for connectives, they do not solve it for negation. E1′ is parsed effectively, but E2′ is not. An atomic attitude under Gibbard's picture, such as "plan to φ," looks clear enough. However, Unwin (taking "plan to φ" as equivalent to "φ is obligatory") shows that there are at least three distinct ways to understand the negation of that attitude:

M) φ is obligatory.
N_1) *A* does not accept that it is obligatory to φ.
N_2) *A* accepts that it is not obligatory to φ.
N_3) *A* accepts that it is obligatory that not-φ. [7]

In different situations, ordinary moral realist-sounding discourse encompasses all three of these types of negation: when I disagree with the claim "φ is right" I can mean any of N_1, N_2, or N_3. So as with the embedding problem, this is an aspect of moral language that expressivists will need to capture. But expressivism looks incapable of symbolizing all three of these negation states. There seem to be only two places to put a not symbol and three places that need them. Whichever way the expressivist decides to symbolize the negation of a pro-attitude towards φ, she will run into problems. If we take the negation

[7] Unwin 2001, 70.

of "plan to φ" to be "plan not to φ," then the expressivist is only able to say that N_3 is a negation of M. Neither N_1 nor N_2 are expressed by "plan not to φ" – N_1 is not a plan about φ at all, and N_2 includes the possibility of planning to φ. We could instead take the negation of "plan to φ" to be "do not plan to φ." This is a weaker statement, since it only encapsulates a lack of plan about φ rather than an active plan against φ. With this interpretation the expressivist can still get N_3, because not planning to φ is compatible with an active plan to not do φ. It also looks like she can get some sort of attitude under which it's permissible but not obligatory to not do φ. But it is unclear whether that attitude would correspond to N_1 or N_2. N_1 and N_2 clearly express different attitudes, and different attitudes will need to be symbolized differently. At this point, it looks like any expressivist semantics is going to have difficulty capturing all three of these senses of negation.

Gibbard's move to plan-based semantics offers hope here, however. Gibbard notes the difference between "rejecting an alternative…and just not choosing it."[8] Viewing the expressivist attitudes as plans offers a way to differentiate more finely than Unwin thinks possible. Here is how Gibbard might put Unwin's three negations:

M′) Plan to φ.
N_1′) *A* does not have a plan regarding φ.
N_2′) *A* plans to do φ or ψ or χ.
N_3′) *A* plans not to φ.

None of these sentences is a negation of M′, strictly speaking. But for Gibbard, since our normative sentences express planning states and are not propositional, the relevant concept to go for is disagreement rather than contradiction. N_1′, N_2′, and N_3′ all express

[8] 2003, 55.

some kind of disagreement with the plan expressed by M′. Each of them expresses this disagreement successfully because each of them is mutually unsatisfiable with planning to φ. Unwin's "φ is obligatory" will require one to have a plan to φ. N_1' expresses a disagreement with having a plan to φ, since someone who holds the attitude N_1' has no plan to φ, having no plan about φ whatsoever. N_2' also expresses a disagreement with having a plan to φ, since it leaves open the possibility of doing other actions instead of φ. And N_3' expresses a disagreement with having a plan to φ, since one will explicitly have a plan to not do φ.

Gibbard's expansion of his planning concept with the idea of a hyperplan is intended to further strengthen this general framework. A hyperplan is a plan which covers all situations and specifies what one will plan to do in each specific one.[9] As we saw in Section 1.1, any normative statement that we give – any plan that we express – can be understood as an illustration of world/plan combinations that are ruled in and out by our commitments. However, statements that rule out *some* worlds in which I have plans that conflict with my commitments or each other are not enough. When I make a normative statement, I need to rule out *all* worlds in which my plans conflict with my commitments. The hyperplan, which covers all plans that I have about everything, enables me to do this. Having a hyperplan provides a descriptivist way to analyze my normative statements as claims about the actual world/hyperplan pair in which I am located. Crucially, the extra weight that the hyperplan provides depends on its being complete; that is, upon its

[9] The claim is not that we actually have a hyperplan. Rather, the planning state I express when I make normative claims is an expression of the hyperplan I take myself to be following and not merely an expression of the local plan I have for the particular situation I am in.

containing a plan for every situation.[10] For a complete hyperplan, φ is either required, φ is impermissible,[11] or φ is permitted (neither required nor prohibited – one may plan to do *p* or to do something else). And the permissibility or impermissibility of an act derives from the completeness of my hyperplan: "It does not follow from a plan's failure to forbid [φ] that the plan permits [φ]. That permission *does* follow from the plan's failure to forbid [φ], plus the completeness of the plan."[12] Only if the hyperplan actually covers all possible cases will I be able to say that φ is permissible.

At this point it again seems that Gibbard has fended off the semantic objection to his view that normative statements express planning states. Gibbard has offered a picture that captures all three senses of negation that Unwin claimed expressivism could not handle.

1.3 – Dreier's Objection to Hyperplans

Unfortunately, Dreier notes that the hyperplanning framework that provides Gibbard with his possible-world semantics still has trouble with Unwin's negation problem. Specifically, Gibbard finds himself on the horns of a dilemma.[13] He can offer his hyperplan view, under which he can get possible-world semantics and very strong license for making inferences like that in Section 1.1 based on which world/plan pairs one can be in. But in this case, he loses the ability to distinguish between N_1' and N_2'. With a complete hyperplan everything that is not forbidden is permitted, but in both N_1' and N_2' φ is not forbidden and thus permitted. So now there is no way for Gibbard to

[10] Gibbard 2003, 56, note 10.
[11] That is, ¬φ is required.
[12] Dreier 2006, 222.
[13] This dilemma is the chief objection raised by Dreier.

differentiate between two attitudes that we think must still differ: one expresses that φ and alternatives are permissible and indifference as to what particular method or combination of methods is chosen, and the other expresses a total lack of a plan regarding φ.[14] Taking this horn thus opens Gibbard back up to Unwin's original charge that expressivism cannot account for all of the actual normative attitudes that people hold. On the other hand, if he wants to keep the distinction in attitudes, Gibbard finds himself on a second horn. If N_1' expresses a total lack of plan about φ, then the hyperplan is not complete, and Gibbard loses the right to use the possible-world semantics he first picked up to address the embedding problem. Without robust possible-world semantics, the planning attitudes we express do not entirely rule out all incompatible world/plan combinations. And if this is the case, then the disjunctive syllogism Gibbard is trying to license to solve the embedding problem holds true only contingently and not necessarily. The difficulty for Gibbard is in how to express an attitude of lack of planning without giving up the completeness of the hyperplan, or how to give up hyperplan completeness without losing the semantics he wants and it is in this problem that Gibbard remains stuck.[15] I offer my solution below.

1.4 – Implicit Planning

I want to effectively take the second horn of Dreier's dilemma, but offer some hope for rescuing Gibbard. I want to accept that in some circumstances it is possible to take oneself to be expressing a total lack of plan about something – an 'incomplete' hyperplan. Since that would be a contradiction in terms, those expressions can instead be

[14] Dreier 2006, 221-2.
[15] Dreier 2006, 227.

understood as a failure to express a hyperplan. However, I also want to say that the nature of planning is such that the 'incomplete' hyperplans that one could take oneself to be expressing with an N_1' statement do not actually exist. In any situation in which one finds oneself, I argue, we do have at least a nebulous plan for action. Only under very rare circumstances do we totally lack a plan, and the very act of being in a situation where we totally lack a plan also makes us form a plan. Then the distinction between N_1' and N_2' can be captured from a third person perspective by saying that someone expressing N_1' is in fact trying to express N_2' but failing to do so. This view will also preserve the use of possible-world semantics when differentiating between N_2' and N_3' and allow the expressivist warrant for making inferences like those in Section 1.1. I need to show how it could be the case both that someone can take themselves to be expressing an N_1' and that there are in fact no such attitudes. I will do this by offering a view of implicit planning.

There is a fact in the world that can clearly differentiate between someone who is undecided (or is expressing a complete hyperplan that permits multiple tied alternatives) and someone who is indifferent (who seems to be expressing an incomplete hyperplan). A person who is indifferent about whether to get a ham sandwich or a tuna sandwich is a person who has never been exposed to a circumstance that requires her to make a plan about what sort of sandwich to get. In the face of any actual set of alternatives, I consider my options and make a plan about what to do. That just is what it is to face a set of actual choices. To realize that one has options and what those options are involves at least a minimal acknowledgment of the pros and cons of each option, and involves at least a minimal judgment either that one option is the best or that some number of options are

equally good. Even if we are speaking only of an imagined situation, I imagine a hypothetical situation by projecting myself into it. My consideration of what I would do may be cursory, and even largely unconscious, but it just is part of the process of modelling a situation. I think the act of planning what to do is implicit in every situation we have ever encountered or imagined. The plans may be created on the fly, they may be largely constructed outside of our awareness by heuristics, but whenever we act or decide, we do so with at least a rudimentary plan implicit in the action or decision that we take.

Note that this view only says that if one is exposed to a situation (in fact or imagination), then one has a plan about the thing to do in that situation. This is to say that exposure to a particular situation is not a necessary condition for the formation of a complete hyperplan, but lack of exposure to a situation *is* a necessary condition for the attempted expression of an incomplete hyperplan.

Consider what happens if Dreier asks me what my lunch preference between a tuna sandwich and a ham sandwich is, and I say "I'm indifferent." What determines if I am failing to express a complete hyperplan or am simply torn between two tied permissible options? We can check (in principle) for previous exposure to a sandwich-choosing situation. Upon discovering that I have never had to choose between ham and tuna before, we can confidently say that my hyperplan is incomplete. Had I ever been exposed to sandwich choice in the past, I would have a plan – poorly thought-out as it might be – about what sandwich to get for lunch. The lack of exposure indicates a lack of specific plan, and thus indicates that I take myself to be expressing an incomplete hyperplan. If instead we discover that I have previously been exposed to sandwich

choice, we now know that I do have at least an implicit plan about how to choose between sandwiches. It is impossible for me to have been exposed to the choice before and not formed some sort of plan, even if the plan is just that all types of sandwich are equally permissible. Thus, I must merely be torn between multiple permissible options. This exposure history and the implicit plans that accompany exposure are facts about the world that we could always determine in principle, and that would always differentiate between someone who has a plan with tied options and someone who has never formed the relevant plan.

This view of implicit planning can now be used to bolster the expressivist case. It is possible that I take myself to be expressing a hyperplan that is incomplete about some situation such as sandwich choice. However, because the creation of a plan is implicit in my exposure to a sandwich choice situation, as soon as I have to choose between sandwiches, I do have a plan for sandwich choice. So if I think my hyperplan is incomplete, I am wrong. For that matter, as soon as I contemplate in the abstract having to choose between sandwiches, I have a plan for sandwich choice. It may be sketchy and largely indifferent, but it is not and cannot be incomplete. And so again, if after contemplating sandwich choice I take my hyperplan to be incomplete, I am wrong. In this sense, it is impossible for any hyperplan I take to be incomplete to ever survive contact with the real world. Being in a situation that would prompt me to express the attitude reflecting an incomplete hyperplan also prompts me to plan. In the very act of expressing an N_1' attitude by saying "I'm indifferent between ham and tuna" while having never had to choose between ham and tuna, I am transforming that N_1' attitude into an N_2' attitude. On my view, it is possible, from a first-person perspective, to make an N_1' claim, but it is

only possible to do so *once*.[16] Furthermore, because of implicit planning, from a third-person perspective it is easy to recognize that when I make the N_1' claim I am actually just failing to correctly express an N_2' claim.

1.5 Final Thoughts on Expressivism

Implicit plans seem to offer a way forward for a Gibbardian expressivist. The view keeps the ability to capture N_1' sentences and provides a fact about the world that determines whether someone is in an N_1' state or an N_2' state. However, because it is only possible to make an N_1' claim once for any given situation, this account may preserve for the expressivist the benefits of hyperplan completeness. Everyday expressivist semantics can essentially behave as if N_1' claims do not exist and be built as though people have complete hyperplans. The use of possible-world semantics will still be warranted, and expressivists will be able to capture the realist-sounding elements of our everyday moral talk. I think understanding implicit planning as a way to a necessary condition of hyperplan incompleteness successfully provides a differentiating fact between N_1 and N_2 negations. This differentiating fact can justify an expressivist in assigning a different attitude to each of the three negation types, while preserving the hyperplan completeness of N_2 and N_3 negations. The fleeting nature of N_1 attitudes means that no problem is posed for hyperplan completeness, and the expressivist can keep possible-world semantics while still being able to capture N_1 attitudes when necessary.

Unfortunately, the requirements on the hyperplan are probably more stringent than the pragmatic view I have advanced above. It is true that due to implicit planning, by the time I have uttered the sentence "I'm indifferent between ham and tuna," I must be

[16] For any given situation.

expressing an N_2' attitude. Nonetheless, this view still implies a moment in which I did not have any plan about sandwich choice, and in order for me to have a hyperplan to express this cannot ever have been the case. While implicit planning may provide a descriptive account of how an individual speaker can take herself to be expressing an N_1' attitude, in order to accomplish the project of justifying the normative force of our moral language, expressivists must have complete hyperplans.[17] And as we have seen, complete hyperplans will rule out the holding – even unconsciously – of N_1' attitudes, let alone attempting to express one. In the end, though my view of implicit plans lets the expressivist push Dreier's dilemma into a relatively small number of cases, the difficulty cannot be overcome. Expressivism simply cannot capture all of the different senses we take ourselves to be using in our moral talk. Accordingly, expressivism should not be taken as a usable metaethical view of the objects of moral language.

2.0 A Consideration of Rational Choice Metaethics

So if expressivism is out, where should the metaethicist turn? One view that perhaps captures some of the naturalistic intuitions that might propel one into expressivism is what I'll call rational choice metaethics. Roughly put, this is the view that the objects of moral language are the choices undertaken by individuals. When we use words like 'right' and 'wrong,' we're referring to the rationality of a particular decision; philosophers being philosophers, 'right' aligns with rational choices and 'wrong' with irrational ones. This view is significantly older than even the emotivism that gave rise to expressivism, being more or less derived from Kant. However, developments in the mathematics and psychology of rational choice theory have turned the view away from

[17] Thanks to Nihar Nilekani for raising a more articulate version of this objection.

abstruse deontology and into naturalistic forms that justify their metaethical claims on the basis of empirically demonstrable facts about human nature. If rational choice metaethics can be made to work, it would offer a promising route for understanding the factors at play in our use of moral language.

The exemplar of modern rational choice metaethics is David Gauthier. His *Morals by Agreement*[18] takes on both the task of explaining how it is that our moral language refers to the rationality of our choices and the project of justifying how it can be possible to derive any first-order moral system from a metaethical picture that claims to speak only of individual interests. I will sketch Gauthier's account of rational choice metaethics and the contractarian first-order view which he thinks is entails, then compare contractarianism with a competing utilitarian view offered by John Harsanyi.[19] I find that Gauthier's way of responding to Harsanyi's justification for utilitarianism undermines rational choice metaethics. Like Gauthier, Harsanyi claims to employ rational choice metaethics. Harsanyi, however, argues that doing so should result in the adoption of utilitarian first-order principles rather than contractarian ones. I will develop an argument to show that if Gauthier's contractarianism can be successfully derived from rational choice metaethics, Harsanyi's utilitarianism can as well. From this basis I will argue that Gauthier himself does not accept his own justification for the claim that rational choices are the objects of our moral language. Rather, Gauthier accepts rational choice metaethics only so long as it appears that contractarianism is the only possible first-order theory that can be derived from rational choice metaethics. When it becomes clear that rational choice metaethics can in fact support multiple first-order theories, Gauthier retreats and

[18] 1986.
[19] In "Morality and the Theory of Rational Behavior," 1977.

claims that rational choice metaethics is justified in terms of its preservation of our starting commitments to respecting individual agency. But since Gauthier's initial claim was that rational choice metaethics can provide the basis for moral language[20] *without* reference to any commitments beyond the axioms of rational choice theory,[21] by making this retreat he implicitly abandons rational choice metaethics altogether. I shall not argue below that rational choice metaethics is therefore unworkable. Rather, I will simply conclude by noting that if the creator of rational choice metaethics does not in fact think it can be justified, it behooves the impartial reader to be suspicious in employing it.

2.1 Gauthier's Rational Choice Metaethics

The first thing to do is to lay out rational choice metaethics (which I will abbreviate as RCME going forward) as Gauthier specifies it. I will then sketch the contractarian moral theory Gauthier derives from RCME to provide an object of contrast for Harsanyi's view. As mentioned, Gauthier's development of RCME constitutes a very small fraction of *Morals by Agreement*, which is mostly devoted to explaining the framework of rational choice theory, justifying the decision to specify choice in terms of preference satisfaction, and attempting to demonstrate that rational choice theory must impose certain constraints on the behavior of any rational actor. Because Gauthier's metaethical discussion is so brief, I will also draw from Harsanyi to clarify the overall RCME view. Nonetheless, RCME as I present it may be incomplete as a metaethical theory. I do think it is possible to offer a characterization of RCME's key features which can distinguish the view from other metaethical systems and provide a basis for understanding why RCME might be appealing.

---

[20] And indeed for moral behavior.
[21] 6.

The first key feature of RCME is its identification of the objects of our moral language with individual goal-directed choices, which is to say rational choices.[22] Gauthier wants to show that when we determine actions to be right or wrong, we are applying the same norms of rationality as in any other situation.[23] Ultimately, when we speak about an action as being right or wrong, we are calling the action rational or irrational, determining its rationality from the axioms and theorems of rational choice theory.[24]

This use of rational choice theory to evaluate actions is the second key feature, and the one that sets RCME apart from other metaethical theories that also identify the objects of moral language with choices.[25] The aim is to show that adhering to moral principles is not a demand of rationality writ large (because if we did not adhere to moral principles, our concepts would become incoherent).[26] Gauthier objects to this route on the grounds that this view of rationality "already includes the moral dimension of impartiality that we seek to generate."[27] Instead, adhering to moral principles is a demand of practical or instrumental rationality. If it is the case that rational choice theory will show cases in which "an individual chooses [instrumentally] rationally only in so far as he constrains his pursuit of his own interest or advantage to conform to principles expressing the impartiality characteristic of morality,"[28] as Gauthier goes on to argue, then morality can be justified purely on the basis of rational self-interest. Moral talk can then be collapsed

---

[22] Harsanyi 42
[23] 2
[24] Harsanyi tends to refer to "decision theory" rather than Gauthier's "rational choice theory," but they are referring to the same field, and for the most part to the same set of principles. Their differences will be discussed below.
[25] Gauthier seems to take Kant as one example of such an alternative view (6)
[26] To grievously mangle Kant's Categorical Imperative.
[27] 6
[28] 4

into talk about the subset of rational choice theory that dictates agents constrain themselves for their own benefit. Seen from this perspective, the RCME project is heavily deflationary. While Gauthier wants to continue to consider moral talk as a special case of talk about rational choices, somewhat distinguished from other rationality talk by its focus on constraint,[29] the claim at the heart of RCME is that moral talk is essentially talk about norms of instrumental rationality.

Gauthier and Harsanyi both (correctly, I think) take RCME to have a number of valuable strengths. First, it provides a clear explanation of whence the normative push in moral talk is derived. Acknowledging that we feel the pressure of instrumental reasons when advancing our own self-interest is easy; if moral reasoning collapses into instrumental reasoning then it is also easy to see where the motivating force is supposed to come from. Second, the view will yield a universal moral standard while respecting individual situations and desires.[30] Because the axioms of rational choice theory – and thus the norms of instrumental reasoning derived from them – are simple and universal, it should follow that any constraining principle that falls out of rational choice theory will be rational for any agent to adopt. At the same time, since the principles of rational choice theory are built to maximize each individual's benefit, following these constraints will still be the best way for an individual agent to get what she wants. Any first-order theory built on RCME will thus be both universal, in that its principles will apply to all agents, and subjective, in that each agent's preferences will determine what the rational choice for that particular agent actually is. Third,[31] RCME offers, or ought to offer, a

[29] 3
[30] Gauthier 6-7
[31] Though Gauthier does not, I believe, call this strength out explicitly.

route by which it can be supported with empirical evidence from human behavior. RCME reduces moral talk to the axioms of rational choice theory (which are empirically supported by their descriptive successes), a set of principles derived therefrom, and the empirically specifiable sets of agents' revealed preferences.[32] Readers are hardly obligated to consider empirical supportability a strength for a metaethical theory, but it certainly can't hurt.

Having laid out RCME, I will now contrast the first-order systems that Gauthier and Harsanyi think follow from it. In so doing, I will demonstrate that if Gauthier's contractarianism successfully follows from RCME, then Harsanyi's utilitarianism does as well.

2.2 Gauthier's Contractarianism

The first element of Gauthier's first-order ethical view to address is his value theory. This will tell us what the aims to be maximized using rational choice theory are and provide the foundation for discussing how maximizing those aims results in the adoption of moral commitments. For Gauthier, value for an individual derives from the satisfaction of the individual's considered revealed preferences: "[o]ne chooses rationally in endeavoring to maximize the fulfillment of those preferences that one holds in a considered way in the choice situation."[33] Because choices are made under conditions of uncertainty, the relevant preferences are preferences for lotteries of various expected outcomes; this move is the foundation of Bayesian decision theory.[34] Under uncertainty,

---

[32] Gauthier 26-8. Revealed preferences are the ends that an agent actually attempts to pursue, thereby 'revealing' what state of affairs she 'prefers' to be the case. Since they are determined from tangible actions, revealed preferences are accessible to observers as well as the agent herself.

[33] 32

[34] Gauthier 44

any coherent preference set will satisfy four axioms: completeness, transitivity, monotonicity, and continuity.[35] Completeness specifies that in any two-element choice situation an agent either prefers one option to the other or is indifferent between them.[36] Transitivity specifies that if an agent prefers A to B and B to C, then the agent prefers A to C.[37] Monotonicity requires that if an agent prefers A to B, then the agent prefers a lottery with the possible outcome A over a lottery with the possible outcome B.[38] Continuity requires that if an agent prefers A to B and B to C that there be one and only one lottery in which A and C are possible outcomes which is indifferent to B.[39] With these four principles, we can "assign utilities measuring [an agent's] preferences over the members of any set of possible outcomes, such that the utility of any lottery taking members of the set as prizes will be its expected utility."[40] The expected utility of any set of possible outcomes and the probability of each outcome (i.e. a lottery) can now be calculated and the expected utilities of different lotteries compared. The value to be maximized using rational choice theory, then, is expected utility, and it can be maximized by choosing the lottery which has the highest expected utility. This should suffice as a sketch of Gauthier's value theory.

The next move is to show that an individual seeking to maximize her expected utility will agree to constraints on her actions in some circumstances; this is the heart of Gauthier's first-order position. Gauthier spends several chapters on the details of this

---

[35] *Ibid*. These are principles commonly used (though not in monolithic agreement) in the literature surrounding preference satisfaction theories of value. Gauthier briefly discusses criticisms of each, but doing so here would be a distraction.
[36] *Ibid*. 39
[37] *Ibid*. 40-1
[38] *Ibid*. 44, and assuming the lotteries are equal in all other respects.
[39] *Ibid*.
[40] *Ibid*.

view, but I am going to give only a very broad overview. Gauthier introduces a basic view of strategic rationality based on the expectation that all participants will behave rationally and that all participants' choices will reflect their knowledge of this expectation.[41] From this framework it can be proven that any situation has at least one set of mutually utility-maximizing strategies in any choice situation involving multiple agents; this is the well-known Nash equilibrium.[42] However, many choice situations have multiple equilibria, or an equilibrium that though rational under strategic choice theory appears to be a suboptimal solution. The Prisoner's Dilemma is the best known example of this type. Here, two prisoners are each given the options of confessing to a crime or remaining silent. If both stay silent, they will each receive minimal sentences. If both confess, they will both receive somewhat longer sentences. And if one confesses and one stays silent, then the blame will be pinned on the silent one, who will serve the maximum sentence while the confessor goes free. Unable to communicate with each other and knowing that the same deal has been offered to both of them, the strategically rational choice for each prisoner is to confess, even though they would have served shorter sentences had they both stayed silent. Hence the equilibrium state of the game is not the optimal state for either participant. Gauthier wants to build a view of strategic rationality that will entail that the optimal choice in such situations can be a viable equilibrium as well, and thus can be selected by strategically rational actors. The aim is to demonstrate a principle of choice that "rationalizes agreement in the way that the principle of expected utility-maximization rationalizes individual choice."[43]

[41] 61. As before, these are basic elements of rational choice theory and will not be further defended here.
[42] *Ibid*. 71
[43] *Ibid*. 118

Gauthier defines cooperation as making "a single joint strategy choice"[44] rather than each individual making their choices in isolation as rational choice theory initially dictates. Of course, under rational choice theory it can be rational to cooperate in choosing a joint strategy if doing so maximizes each individual's expected utility. This choice of joint strategy can be accomplished through a bargaining process in which each participant follows the principle of minimax relative concession. As Gauthier sketches it, minimax relative concession[45] is the~~This boils down~~[46] ~~to a~~ principle that each participant can rationally accept a joint strategy if it does not require her to concede more than anyone else from her initial bargaining position. If no party has lost more than any other by entering into the agreement, this is a justification for all of the parties to consider the bargain fair.~~position.~~ If there is an optimal equilibrium of joint strategies for approaching a given choice situation, Gauthier claims, a bargaining process using this principle will identify it and it will be rational for all participants to agree to it. [47] The participants in a Prisoner's Dilemma, for instance, can agree on a joint strategy of mutually clamming up by bargaining over the four possible joint strategies that they could adopt. Because mutual silence requires each to concede the same risk compared to their initial positions of confessing and the benefit of adopting the strategy of mutual silence exceeds that of mutually confessing, they can rationally agree to keep silent, breaking out of the undesirable equilibrium and reaching the optimal outcome of the choice situation.

---

[44] *Ibid*. 120
[45] There is math, or something like math, but I will pass over it in silence.
[46] ~~There is math, or something like math, but I will pass over it in silence.~~
[47] Gauthier 133-40

However, this process does not explain why it is rational to hold oneself to any agreement made in a joint strategy bargain when one could act on an individual strategy with higher expected utility than the agreed joint strategy.[48] To address this point Gauthier introduces the concept of constrained maximization. Because an optimal joint strategy for a choice situation is one in which each participant receives the maximum utility compatible with the other participants,[49] it can be said that each agent has rationally chosen to constrain their actions based on the utilities their actions will afford others. A constrained maximizer is thus an agent who "seeks in some situations to maximize her utility, given not the strategies but the utilities of those with whom she interacts."[50] This can be understood as a conditional disposition to follow a joint strategy given that doing so would leave her at least as well off as she would be if everyone pursued an individual strategy and that others have this disposition as well.[51] This second element is meant to make it irrational to agree to joint strategies and then choose not to follow through on them: if an agent does so, others who have the conditional disposition to cooperate will stop cooperating with her, and she will lose out on both the benefits of cooperating and the benefits of agreeing to cooperate and then defecting.[52]

Gauthier thinks that this chain of reasoning shows that the conditional disposition to cooperate and the minimax relative concession process for cooperation can be proven rational from only the basic principles of rational choice theory. For the purpose of this discussion, it is unnecessary to consider whether or not he is correct. Instead, I will move

---

[48] *Ibid*. 166. This is Hobbes' Foole problem, which Gauthier discusses at length. I will not recapitulate the entire chapter here, but only give a summary of what Gauthier thinks to be the solution.
[49] Recall that the process of minimax relative concession bargaining identifies the equilibria that satisfy this definition of optimality.
[50] Gauthier 167
[51] *Ibid*.
[52] *Ibid*. 171-2

on to detail Harsanyi's utilitarian account in order to build the argument that both views can follow from RCME.

2.3 Harsanyi's Utilitarianism

Harsanyi offers two different proofs that rule utilitarianism follows from adoption of RCME; I will only be considering the second, which explicitly lays out axioms of rational choice theory and demonstrates how consideration of overall utility can be derived from them. Harsanyi begins by defining a difference between moral preferences and other preferences. Moral preferences are the preferences we have about how society is to be structured under the assumption that we do not know our position in that society.[53] This ignorance (which is only theoretical) allows for impartial comparison between different societal arrangements. Harsanyi will attempt to show that RCME entails that maximizing the fulfillment of one's moral preferences – that is, one's individual expected utility in choice situations about societal rules – is necessarily best served by maximizing overall utility in those choice situations.

Harsanyi begins by introducing three axioms of rational choice theory. The first (A1) states that the personal preferences of all individuals satisfy the Bayesian rationality postulates.[54] Second (A2), at least one individual's (referred to below as individual *i*) moral preferences satisfy the Bayesian rationality postulates. This is just to say that I follow the same standards of rationality in determining how to pursue societal interests as I do in pursuing my own.[55] Third (A3), if at least one individual (referred to below as *j*) prefers A

---

[53] Harsanyi 43
[54] 48, note 17. Two of these principles correspond to Gauthier's completeness and continuity (see p. 23 above). The other two are actually weaker than Gauthier's, though his transitivity and monotonicity requirements can be specified from within Harsanyi's framework.
[55] Harsanyi 48

to B and no individual prefers B to A, then an individual satisfying A1 and A2 will morally prefer A to B.[56] A3 is a statement of Pareto efficiency that in practice means that if *j*'s preference can be satisfied without thereby thwarting any other individual's preference, then satisfying that preference is more efficient than not satisfying that preference and *i* will prefer to satisfy *j*'s preference. It follows from A1 that each individual's (for instance, *j*'s) personal preference set can be represented by a von Neumann-Morgenstern utility function U_j and from A2 that *i*'s moral preference set can be similarly specified as W_i.[57] By adding in A3 we can see that W_i must be equivalent to the expected personal utilities of all other individuals; Harsanyi calls this Theorem T.[58] Finally, he adds a fourth axiom (A4), symmetry. This specifies that T is symmetric, meaning it can be computed the same way for any individual *i* and set of other individuals $j = 1, \dots, n$, regardless of how many *j*'s there actually are.[59] Under these conditions, "a rational individual will always choose that particular social system that would maximize his expected utility, that is, the quantity…representing the arithmetic mean of all individual utility levels in society."[60] In other words, given that I am Bayesian rational in both my personal and moral preferences, that my moral preferences are Pareto efficient, and that a social welfare function can be computed identically for any individual's moral preferences, including my own, I will successfully maximize my own expected utility around my own personal preference set[61] by maximizing overall utility.

---

[56] *Ibid*.
[57] *Ibid*. 49
[58] *Ibid*. Note that because A2 imposes coherence requirements on W_i, the personal preferences of the other individuals must be compatible. T thus rules out options that would be unacceptable to any particular individual out of any social utility function.
[59] *Ibid*.
[60] *Ibid*. 46 for the quote and 49 for the claim that it follows from A1-A4 and T.
[61] Regardless of what that preference set contains.

Having laid out two first-order normative views that are alleged to follow from RCME, I will now address Gauthier's objection to Harsanyi. The aim, again, will be to prove that if Gauthier's contractarianism follows from the postulates of RCME, then Harsanyi's contractarianism does as well. Once that argument is complete, we will return to the metaethical level to review Gauthier's abandonment of RCME.

2.4 If One, Then Both

I will proceed by showing that Gauthier ought to accept A1-A4 and T given his own claims about the nature of rational choice theory. A1 is a basic postulate of rational choice theory, and one which Gauthier has himself imposed in his discussion of preferences.[62] Gauthier states that rational choice theory is agnostic about the source and contents of our preferences,[63] because if the theory dictated which preferences agents can hold by means of anything other than his four coherence conditions it would be importing external moral assumptions. Preferences about how societies are structured are preferences, so the coherence requirements of A1, which Gauthier accepts, will also apply to Harsanyi's moral preferences.[64] Accordingly, Gauthier must accept A2. Since Gauthier and Harsanyi agree that RCME only refers to an individual's own preferences, Gauthier has to at least accept A3 insofar as agents are indifferent between options that do not affect their own expected utilities. Whether we in fact have moral preferences for

---

[62] 44

[63] 22: "The theory does not analyze particular relations of preference."

[64] The objection can be made that since Harsanyi's definition of moral preferences already builds in a level of ignorance of one's current position, Gauthier could reject A2 on the grounds that it's impossible and/or not necessary to hold moral preferences. (Thanks to Stan Husi for raising this issue.) I think this objection holds against the first argument Harsanyi gives in "Morality and the Theory of Rational Behavior" but not against the axiomatic argument being covered here; for the axiomatic argument, moral preferences can be defined just as the preferences one has for the structure of society, with the impartiality component removed. As will be shown, in the axiomatic argument the impartiality is intended to derive from A4. With moral preference understood in this weaker sense, Gauthier has no reason to reject A2.

Pareto efficiency in the way Harsanyi claims is then an empirical question, but it seems plausible that we do, and the claim that we do can be made with reference only to A1 and A2. Therefore Gauthier should accept A3, or at least not reject it on theoretical grounds. Unlike A1-A3, Gauthier does not use social welfare functions, but since T can be derived only from A1-A3, which he accepts, Gauthier must also accept that an individual's moral preferences can be specified by T. The point of departure is A4.

Gauthier denies A4 on the grounds that individual utility functions can't be proven to be symmetric because individual utility functions are not strictly comparable. According to Gauthier, when we calculate a hypothetical individual utility function for an individual of unknown position, we substitute our conception of what our own preferences in that unknown position would be for the preferences that the hypothesized individual would actually have.[65] Attempts to impartially compute an individual utility function for anyone other than oneself will therefore necessarily fail. And since different individuals' utility functions must be derived impartially to be comparable in a way that permits symmetry, symmetry is impossible. However, I find that Harsanyi successfully addresses this worry. He explicitly states that part of impartiality is at least the attempt not to do what Gauthier takes us to be doing in computing hypothetical individual utility functions.[66] We might not do it well, but to uphold the instrumental norm of impartiality,[67] we ought to try to project what the hypothetical agent's preferences would be rather than using our own. Harsanyi also offers a further reply by introducing a similarity postulate. This is simply the view that it is reasonable, if not strictly rational, to

[65] 240-1
[66] Harsanyi 50
[67] This is only an instrumental norm in this case, and so doesn't violate Gauthier's requirement of not importing moral principles.

assume that once all empirically demonstrable differences between myself and the hypothesized individual are accounted for "our basic psychological reactions to any given alternative will be much the same."[68] Maintaining an epistemic concern that there may still be some "hidden and unobservable differences in [our] psychological feelings" would be unwarranted.[69] The point is that, while interpersonal utility comparison is hardly exact in practice, there is nothing that makes it impossible in principle. Furthermore, in our actual attempts to compare our situations with another's, we do attempt to follow the norm of impartiality in the way that Harsanyi describes. In sum, if we do fail at interpersonal utility comparison, it is not a *necessary* failure, and thus does not support a rejection of A4. Accordingly, Gauthier's objection to A4 fails and he ought to accept A4.

But if Gauthier, beginning from the position of only using the axioms of rational choice theory, would accept A1-A4 and T, then RCME equally justifies contractarianism and rule utilitarianism. In fact, Gauthier himself acknowledges that if interpersonal utility comparison is possible, then utilitarianism can follow from RCME.[70] I think this demonstration impels Gauthier to turn away from his foundational commitment not to employ any principles outside of those of rational choice theory, with seriously problematic consequences for RCME.

2.5 Gauthier Abandons Rational Choice Metaethics

So what reason do we have to choose contractarianism over utilitarianism if both can follow from RCME? According to Gauthier, the distinction is that contractarianism

---

[68] Harsanyi 50
[69] *Ibid*.
[70] 127

captures the active involvement of the participants and the requirement that joint strategy adoption be voluntary (128). But the volitional status of the agents is not part of rational choice theory, which is completely agnostic about the internal lives of agents. Recall that Gauthier needs to specify agents' preferences in terms of revealed preferences, which are read externally from agents' behavior and not internally from the agent's sentiments, and which therefore do not make reference to agents' volition. If volition is not a necessary element of rational choice theory, then it can't be a requirement of RCME either. The aim in attempting to build RCME was to get a metaethics that generates first-order moral principles without building in *any* of them at the outset. Gauthier explicitly states at the start of the project that for RCME to succeed, it must generate reasons for constraint on action "strictly as rational principles for choice, and so without introducing prior moral assumptions."[71] So the question of volition that Gauthier wants to raise is ultimately immaterial to whether any particular first-order theory follows from RCME. Harsanyi's utilitarianism still follows just as much as contractarianism does. While utilitarianism's disregard for volition might in fact be a reason to prefer some other view, it's not a distinctively metaethical reason to do so unless principles outside the scope of RCME are in play.

If contractarianism is right and utilitarianism is wrong, as Gauthier continues to insist, it can't be because utilitarianism fails to follow from RCME. So instead it must be because utilitarianism fails to follow from some other set of foundational metaethical commitments. Gauthier does not say what these other commitments might be or why they ought to be adopted. RCME's principles are explicitly exhausted by the identification of

[71] 6

the object of moral language with choices and the specification that rational choice theory is to be employed for deriving first-order principles about the rationality or irrationality of those choices. So whatever Gauthier's other metaethical commitments are, they can't be part of RCME; I am confident that this would be even more clear if Gauthier ever got around to specifying those other commitments. Though he is not explicit, one might say that his revealed preference is for a view that builds in some degree of moral requirements around volition and/or autonomy.[72] The bottom line is that whatever Gauthier's metaethical theory is, it's not RCME when his first-order views are on the line.

2.6 Final Considerations on Rational Choice Metaethics

~~This is a twisted and dissatisfying line of argument with a fairly weak conclusion.~~ Gauthier's abandonment of RCME under fire doesn't prove that it's wrong. Nor does abandonment of RCME show that contractarianism is wrong. The knock-down argument against Gauthier is his inability to address the Foole problem,[73] which is not a distinctively metaethical line of attack and is therefore beyond the scope of this work. However, I think that I have shown that Gauthier thinks RCME is incapable of capturing some of our common moral intuitions, specifically those centered on the volition and autonomy of agents. RCME was supposed to yield a ground for all of our moral intuitions. If it must resort to principles beyond those of rational choice theory to capture some of those intuitions, then that project has not succeeded. If one has those other

[72] One might also say that he is throwing an ad hoc response at Harsanyi out of a refusal to accept that RCME justifies utilitarianism, but that claim is less charitable and less productive.
[73] Specifically in 175-8, wherein by trying to prove that an unconstrained maximizer will usually receive very little benefit from defecting, he concomitantly proves that an unconstrained maximizer will always receive a nonzero benefit from defecting, will always know that there is a nonzero benefit to defecting, and therefore will always be rational in doing so.

intuitional commitments, there is no choice but to conclude that first~~the project has failed. First~~-order normative theories cannot be successfully built on nothing but rational choice theory. To return more squarely to metathics, RCME's failure does not prove that choices are not the objects of our moral language. However, it does show that the standard for evaluating those choices cannot be rational choice theory if we are to derive first-order views consistent with our moral intuitions. Absent one of its defining features, RCME does not succeed.

3.0 Overall Conclusions

I think the most significant lesson to be learned from this foray into metaethical failure is a methodological one. Expressivism and RCME present two cases of metaethical theories that founder based on their inability to justify some moral intuition/aspect of moral language that we take ourselves to experience and practice. Expressivism fails because it cannot capture all the senses in which we can consider a moral claim negated. RCME fails because it does not capture our sense that active participation is necessary for choice in general and especially for moral choice. Perhaps, then, the most useful way forward in metaethics is to actually take the route that Gauthier claims to, and accept that "there will be differences, perhaps significant, between the…constraints supported by our argument, and the morality learned from parents and peers, priests and teachers."[74] In other words, perhaps the best use of metaethics is not to justify the moral intuitions we hold, or the moral language we use, but rather to convincingly argue that only certain moral language actually refers to anything. If instead

[74] 6

we remain committed to the view that metaethical theories that cannot explain everything we say necessarily fail, I fear no successful metaethical view can ever be found.

Bibliography

Dreier, James. "Negation for Expressivists: A Collection of Problems with a Suggestion for their Solution." *Oxford Studies in Metaethics*, 2006: 217-33.

Gauthier, David. *Morals by Agreement*. New York: Oxford University Press, 1986.

Gibbard, Allan. *Thinking How to Live.* Cambridge: Harvard University Press, 2003.

—. *Wise Choices, Apt Feelings.* Cambridge: Harvard University Press, 1990.

Harsanyi, John C. "Morality and the Theory of Rational Behaviour." *Social Research*, 1977: 39-62.

Joyce, Richard. "Moral Anti-Realism." *Stanford Encyclopedia of Philosophy.* June 21, 2009. http://plato.stanford.edu/archives/sum2009/entries/moral-anti-realism (accessed November 2, 2014).

Roojen, Mark van. "Expressivism and Irrationality." *The Philosophical Review*, 1996: 311-35.

Sayre-McCord, Geoff. "Moral Realism." *Stanford Encyclopedia of Philosophy.* June 21, 2011. http://plato.stanford.edu/archives/sum2011/entries/moral-realism (accessed November 2, 2014).

Schroeder, Mark. *Noncognitivism in Ethics.* New York: Routledge, 2010.

Unwin, Nicholas. "Norms and Negation: A Problem for Gibbard's Logic." *The Philosophical Quarterly*, 2001: 60-75.

Unwin, Nicholas. "Quasi-Realism, Negation, and the Frege-Geach Problem." *The Philosophical Quarterly*, 1999: 337-52.