Journal of Ethics and Social Philosophy
VOL. 17, NO. 3 · JUNE 2020
https://doi.org/10.26556/jesp.v17i3.792

THE ELIGIBILITY OF RULE UTILITARIANISM

David Mokriski

THE ELIGIBILITY THEORY OF MEANING (ETM), also known as the doctrine of reference magnetism, has played a significant role in recent discussions of metaphysics and philosophy more generally, including metaethics. According to ETM, the referent of a predicate is the property that best balances fit-with-usage, or "charity," and eligibility, where eligibility is a function of how metaphysically natural the property is. ETM is motivated by its ability to avoid intolerable levels of semantic indeterminacy and secure shared reference between disputing parties. This sort of metasemantics has the potential to be friendly toward somewhat revisionary theories—theories that do not fit so well with some of our considered judgments—since the superior naturalness of a candidate referent can outweigh some mismatch with usage. In this way, highly natural properties act as "reference magnets," securing our reference despite apparent counterexamples and otherwise less than optimal fit.

Using considerations of naturalness and eligibility, several philosophers have recently argued for somewhat revisionary theories in epistemology, ontology, and the metaphysics of truth.[1] In this paper, I add a similar argument to the stock, applying these considerations to normative ethics. In particular, I argue that the theory of rule utilitarianism (RU) achieves a high balance of charity and eligibility. I will not argue that it achieves the *best* balance, relative to all possible (or popular) ethical theories, for that would be too ambitious. However, by comparing RU to two of its common rivals, act utilitarianism (AU) and Rossian pluralism (RP), I show how RU strikes a good balance between two extremes. On the one hand, AU achieves a high degree of eligibility but only at a significant cost of charity, while RP does the opposite, fitting very nicely with our consid-

1 Weatherson ("What Good Are Counterexamples?") defends the "justified true belief" theory of knowledge against Gettier's apparent counterexamples in Gettier, "Is Justified True Belief Knowledge?" Sider defends ontologists' ability to reject "common sense" views about what objects exist ("Ontological Realism"), and Edwards defends "representational" theories of truth against apparent counterexamples that threaten their scope ("Naturalness, Representation, and the Metaphysics of Truth").

ered judgments but at the price of low eligibility. A compromise between these factors would be preferable, and RU fits the bill, making it a promising theory.

My plan for the paper is as follows. In section 1, I introduce and motivate ETM and give a brief overview of metaphysical naturalness. In section 2, I give a rough account of the three rival theories in normative ethics that I will be comparing on grounds of charity and eligibility. In section 3, I take as my starting point the idea that we have some sort of a moral reason to "promote the good" and introduce five questions that must be addressed in order to clarify and precisify this thought. Each of these questions represents a dilemma for the theorist who endorses ETM, for one answer leads to a far more eligible theory like AU, while the other leads to a much more charitable one like RP. I then show how RU sidesteps each dilemma, achieving a high degree of charity without sacrificing much eligibility. In section 4, I address some objections regarding whether RU is really as charitable as I claim. Finally, I conclude in section 5 with a brief discussion of the metaphilosophical costs of denying ETM or of downplaying the role of eligibility in metaethics.

1. ELIGIBILITY AND NATURALNESS

Meaning is not just a function of use. At least, this is the lesson that Lewis draws from Putnam's "model-theoretic argument" against metaphysical realism, Kripke's Wittgenstein-inspired semantic skepticism, and similar puzzles.[2] For any given term (e.g., "green"), there are far too many candidate referents that fit equally well with our usage (e.g., *being green*, *being grue*).[3] Likewise, for some terms (e.g., "gold"), there are bizarre candidates that may fit *better* with our usage

2 Lewis, "New Work for a Theory of Universals" and "Putnam's Paradox"; Putnam, "Realism and Reason"; Kripke, *Wittgenstein on Rules and Private Language*.

3 The predicate "grue" originated in Goodman, *Fact, Fiction, and Forecast*. An object is grue iff it is either green and discovered before some arbitrary future date (say, AD 3000) or blue and not so discovered. One might complain that *being grue* does *not* in fact fit with our usage of "green," since presumably fit-with-usage includes future and counterfactual usage as well as past and actual usage. Since, if prompted, we would probably say things like, "Green is the color of *all* emeralds, not only ones discovered before AD 3000," this may seem to disqualify *being grue* from being the referent of "green" on usage grounds alone. However, the problem is that there is a possible referent for "emerald," namely *gremerald*—where an object is a gremerald iff it is either an emerald and discovered before AD 3000 or a sapphire and not so discovered—and an interpretation that assigns *grue* to "green" and *gremerald* to "emerald" will fit with our usage as well as the intuitively intended one, at least when it comes to the linguistic dispositions mentioned so far. In general, it is always possible to take the intuitively intended interpretation and perform systematic permutations that yield bizarre (and intuitively unintended) interpretations that nevertheless fit with our usage equally well.

(e.g., *being gold-or-fool's-gold*) than their intuitively "intended" rivals (e.g., *being gold*).[4] As a proposed solution to these puzzles, ETM holds that reference is determined by *two* factors: how well a candidate referent fits with our usage of the term in question and the nature of the candidate referent itself.[5] In short, the referent of a predicate is the property that best balances the twin constraints of charity and eligibility.[6]

In addition to resolving indeterminacy, ETM has been put to work in securing shared reference between disputing parties, thus explaining the possibility of genuine disagreement despite diverging usage of a common term.[7] In metaethics, ETM has been proposed as a solution to the Moral Twin Earth challenge, offering an explanation of how our core moral term (e.g., "morally permissible") and the orthographically identical moral term of our "twins" on Twin Earth could refer to the same property, even if our patterns of usage are somewhat different (e.g., we are committed deontologists and our twins committed consequentialists).[8] If there is one highly natural property in the vicinity, then it would serve as a "reference magnet" and secure shared reference despite our somewhat different usages.[9] The merit of this response lies in the fact that it follows from an independently plausible, *general* metasemantics, solving a problem in metaethics that has plagued some versions of moral realism for decades.[10]

It is worth here clarifying the metaphysics that is presupposed by ETM, namely the distinction between natural and unnatural properties. Metaphysical natu-

4 Sider, "Criteria of Personal Identity and the Limits of Conceptual Analysis," 191.

5 For the most comprehensive discussions of ETM, see Lewis, "New Work for a Theory of Universals," 370, and "Putnam's Paradox"; and Sider, *Writing the Book of the World*, sec. 3.2. For arguments that the eligibility constraint is independently motivated and hence no mere *ad hoc* solution to the semantic puzzles, see Williams, "Eligibility and Inscrutability," 371; and Sider, *Writing the Book of the World*, 28.

6 Weatherson argues that this statement of ETM is an oversimplification but still a useful heuristic ("The Role of Naturalness in Lewis's Theory of Meaning"). For some problems with the simple account and suggestions on how to supplement it, see Williams, "Eligibility and Inscrutability"; and Hawthorne, "Craziness and Metasemantics."

7 Weatherson, "What Good Are Counterexamples?" 7–8; Sider, "Ontological Realism," sec. 11.

8 For the problem, see Horgan and Timmons, "New Wave Moral Realism Meets Moral Twin Earth." For the solution that appeals to ETM, see van Roojen, "Knowing Enough to Disagree"; Edwards, "The Eligibility of Ethical Naturalism"; and Dunaway and McPherson, "Reference Magnetism as a Solution to the Moral Twin Earth Problem."

9 In other words, even though distinct properties fit best with each community's respective usage, the same highly natural property would achieve the best *balance* of fit-with-usage and naturalness for both.

10 Dunaway and McPherson, "Reference Magnetism as a Solution to the Moral Twin Earth Problem," 641.

ralness is the gradation of properties exemplified by the pair *being green* and *being grue*; the property *being green* is more natural—less "gerrymandered"—than the property *being grue*.[11] The naturalness dimension ranges from the *perfectly natural* to the hopelessly gruesome. The perfectly natural properties are the fundamental ones, and this binary distinction of perfect naturalness is used to analyze the scalar notion of comparative naturalness on the traditional Lewisian view.[12] Every (less than perfectly natural) property has a canonical definition—a definition in terms of the perfectly natural properties (and logical operators)—and one property is more natural than another to the extent that the canonical definition of the former is less gerrymandered than that of the latter. Factors that contribute to gerrymanderedness include length, complexity, and the miscellaneousness of its constituents.[13] For example, the canonical definition of *being green* is intuitively much less gerrymandered than that of *being grue* (e.g., *being green-and-discovered-before-AD-3000-or-blue-and-not-so-discovered*); the latter is longer and more complex, and its constituents are more miscellaneous. Note that this difference in gerrymanderedness is highly plausible even given our inability to produce the full canonical definitions of such properties—we need not know the basis of *being green* in fundamental reality in order to know that it is a more natural property than *being grue*. We will return to this point soon.

Many theorists, even those otherwise sympathetic toward the appeal to naturalness in philosophy, have rejected the traditional Lewisian account of comparative naturalness, typically in favor of primitive degrees of naturalness.[14] However, I think many of the standard objections are overstated and the traditional account is worth maintaining. First, the traditional account is reductive, analyzing comparative naturalness in terms of fundamentality, a notion that many already have use for in philosophy. Second, the traditional account captures the paradigm examples of differences in comparative naturalness very well (e.g., *green* versus *grue*)—unnatural properties always seem like "merely arbitrary constructions" compared to their more natural counterparts.[15] Third, the traditional account does not leave our comparative naturalness judgments

11 For the most systematic discussions of metaphysical naturalness, see Lewis, "New Work for a Theory of Universals"; Sider, *Writing the Book of the World*; and Dorr and Hawthorne, "Naturalness."

12 Lewis, "Putnam's Paradox," 228.

13 Guigon, "Overall Similarity, Natural Properties, and Paraphrases," 8.

14 For objections, see Williams, "Eligibility and Inscrutability"; and Hawthorne, "Craziness and Metasemantics." Dunaway and McPherson are among those who opt for primitive degrees of naturalness ("Reference Magnetism as a Solution to the Moral Twin Earth Problem").

15 Hirsch, *Dividing Reality*, 55.

unconstrained the way primitive degrees of naturalness seem to.[16] On the primitivist view, different theorists are bound to find their preferred properties to be more natural than those of their rivals, whereas on the traditional view, there is pressure to converge on naturalness judgments insofar as our judgments of gerrymanderedness converge.

In spite of these benefits, many have objected that the traditional view is inadequate. The main objection is the worry that many properties of interest only have canonical definitions that are infinitely long, and there is no way to distinguish the comparative naturalness of such properties on the traditional view.[17] However, as Guigon notes, this objection ignores the fact that length is not the only factor that contributes to gerrymanderedness.[18] Even if two properties have canonical definitions that are both infinitely long, one may be more complex (e.g., be a conjunction of disjunctions rather than just an extended disjunction) or one may have more miscellaneous constituents. Furthermore, this objection is predicated on the idea that infinite canonical definitions of interesting properties are common. However, this seems to assume a sort of hyper-microphysicalist view according to which the only perfectly natural properties are certain microphysical ones, and all other properties are only definable as infinite disjunctions of realizations in microphysical terms. But why should we think the only canonical definition of, say, *being a person* is of the form *being* P_1 *or* P_2 *or* …, where each P_i is a complete description of a possible person in microphysical terms? As Sider discusses, many of these properties are more plausibly defined in finite functional terms.[19] Likewise, we may have good reasons to countenance some properties at levels other than just the microphysical as perfectly natural.[20]

Even if one grants the traditional Lewisian view is correct, one might complain that it leaves us completely in the dark about the comparative naturalness facts, since we rarely know any canonical definitions in full detail. However, as

16 Dunaway and McPherson, "Reference Magnetism as a Solution to the Moral Twin Earth Problem," 653.

17 Another common objection is the worry that the traditional view yields the wrong verdict about "reasonably" natural properties like *being green* or *being a person*, since such properties are plausibly wildly complex when spelled out in terms of the fundamental properties. However, we must keep in mind that naturalness is a *comparative* matter, so when we think of properties like *being green* or *being a person* as "reasonably" natural, this is because they are far simpler than surrounding properties like *being grue* or *being a person-not-born-on-a-Tuesday*. It does not matter that such "reasonably" natural properties are much more complex, and hence much less natural, than fundamental properties like *spin* or *mass*.

18 Guigon, "Overall Similarity, Natural Properties, and Paraphrases," 7–9.

19 Sider, *Writing the Book of the World*, 130.

20 For discussion, see Schaffer, "Two Conceptions of Sparse Properties."

I mentioned above, we can typically gauge the gerrymanderedness of such canonical definitions even when we only have access to the ordinary-language definitions. For instance, we can tell that the canonical definition of *being red* is less gerrymandered than that of *being green-and-discovered-before-AD-3000-or-blue-and-not-so-discovered*, even without having access to either of these canonical definitions. This is because the canonical definition of *being red*, however gerrymandered it might be, is intuitively on par with those of *being green* and *being blue*, and the former is therefore much less gerrymandered than a complex and arbitrary construction involving both of the latter as well as the further complications involving *being discovered*, *AD 3000*, and the Boolean operators.

This suggests the following epistemology of comparative naturalness: to the extent that the ordinary-language definition of a property is more complex and arbitrary than another, this is evidence that the former property is less natural than the latter.[21] After all, if most of the terms that occur in two such ordinary-language definitions denote properties that are roughly on par with respect to naturalness, then any difference in the gerrymanderedness of these two definitions will roughly track a difference in the gerrymanderedness of their corresponding canonical definitions. In general, the more gerrymandered the ordinary-language definition of a property is, the more gerrymandered its canonical definition will be, since the canonical definition is typically obtained by taking the ordinary-language definition and replacing its terms with the canonical definitions of their referents. For example, the canonical definition of *being gricular* (i.e., *being green or circular*) is plausibly "*being G or C*," where *G* and *C* are the canonical definitions of *being green* and *being circular*, respectively.

In my argument for RU, I will appeal to this methodology quite a bit, gauging the naturalness of a property by how gerrymandered its ordinary-language definition is. The definition of $permissible_{AU}$ is highly simple and nonarbitrary, but AU does severe damage to our moral intuitions. RP, on the other hand, fits very nicely with our moral intuitions, but the definition of $permissible_{RP}$ ends up being highly complex and arbitrary. Compared to AU and RP, RU achieves a nice balance; the definition of $permissible_{RU}$ is moderately simple and nonarbitrary, while achieving a moderately high degree of fit with our considered judgments. I will soon make the case for this in greater detail.

It is worth briefly pausing here to outline the package of assumptions about naturalness that I made in this section and on which the main argument of this

21 This methodology fails when our language is of the bizarre sort discussed in Hirsch (*Dividing Reality*), in which there are primitive terms like "gricular" for unnatural properties like *being green or circular* and complex expressions for more natural properties like *being green*. However, ordinary languages are typically not like this.

paper depends.[22] While I think there is good motivation for each, I also grant that they are highly controversial, and opponents of my conclusion would not be in bad company if they ended up denying them. These assumptions are:

- N1. The comparative naturalness of a property is a function of how gerrymandered its canonical definition is.
- N2. The degree of gerrymanderedness of such a definition is a function of length but also other factors such as complexity and the miscellaneousness of its constituents.
- N3. The perfectly natural properties that figure in such canonical definitions are not necessarily limited to microphysical properties.
- N4. The degree of gerrymanderedness of a property's ordinary-language definition is (typically) a good guide to the degree of gerrymanderedness of that property's canonical definition.

Theorists who think we should ultimately reject one or more of these assumptions may still find it valuable to see their implications for normative theory on a metaethics that includes ETM. Before I defend these implications, I will give a brief, and somewhat rough, overview of each theory I will be comparing.

2. THREE RIVALS IN NORMATIVE ETHICS

AU, RP, and RU obviously do not exhaust the options in normative ethics. Yet they make for useful comparisons, especially as different ways of developing a theory of normative ethics from the plausible starting point that, other things being equal, it is in some sense morally preferable to make the world better for all.[23] I will interpret each view as primarily a theory of moral *permissibility*, though much of what I say could instead be put in terms of moral *rightness*, what we morally *should* or *ought* to do, or what we have most moral *reason* to do.

Concerning metaethics, I will assume a naturalist or reductive view, according to which moral properties are identical to naturalistic properties of the sort that are in principle investigable by natural science (e.g., *being an action that maximizes overall happiness*).[24] I make this assumption because on the non-re-

22 Thanks to an anonymous reviewer for this suggestion.

23 Hooker, for instance, often defends his theory of rule consequentialism by making comparisons with RP, as well as with act consequentialism ("Ross-Style Pluralism versus Rule-Consequentialism"; *Ideal Code, Real World*). Similarly, Sidgwick's discussion of the "methods" of ethics involves a systematic comparison of utilitarianism (and egoism) with RP, using the label "intuitionism" for the latter (*Methods of Ethics*).

24 Naturalists include Railton, "Moral Realism"; Boyd, "How to Be a Moral Realist"; and Brink, *Moral Realism and the Foundation of Ethics*.

ductive view, on which moral properties are held to be fundamental, there is no difference in eligibility between AU, RU, RP, or any other first-order ethical theory; on all such theories, the moral property is perfectly natural, no matter how unnatural the naturalistic property is on which the purported fundamental moral property supervenes. Furthermore, I assume that the naturalness of a moral property that is identical to naturalistic property N is determined by the canonical definition of N. This rules out the possibility of a naturalistic property with a highly gerrymandered canonical definition turning out to be highly natural simply because it is identical to a moral property.[25]

Finally, I also assume that the reductive view in question will identify the moral property with the naturalistic property that is described by the substantive first-order theory (e.g., *being an action that maximizes utility, respects autonomy*) rather than some sort of a response-dependent construction (e.g., *being permitted by the moral framework that an idealized subject would endorse*).[26] On such a response-dependent view, the naturalness of the moral property would be settled by the canonical definition of this response-dependent property itself and would be independent of the content of the first-order theory it happens to track, and so all first-order theories would once again be tied for eligibility. As a matter of fact, given my argument in this paper, I think the prospects for maintaining the first-order theory of RP are most promising if we accept a response-dependent reduction of the sort described above. This could give us the plausible first-order consequences of RP without its otherwise poor degree of eligibility.

2.1. Act Utilitarianism

According to one characterization of AU, an action is morally permissible if and only if it leads to at least as much overall well-being as any other available action.[27] Well-being is typically characterized as the balance of pleasure over pain, preference satisfaction, or else some combination of features that includes these as well as things like autonomy, friendship, accomplishments, etc.[28] Although

25 Thanks to an anonymous reviewer on this point.

26 Thanks to an anonymous reviewer for encouraging me to make this assumption explicit.

27 Proponents of AU include Bentham, *Introduction to the Principles of Morals and Legislation*; Mill, "Utilitarianism"; Sidgwick, *Methods of Ethics*; Singer, "Is Act-Utilitarianism Self-Defeating?"; Smart, "An Outline of a System of Utilitarian Ethics"; and, more controversially, Hare, *Moral Thinking*. It may be unfair to characterize some of these earlier authors as proponents of AU over RU, given that this distinction was not made explicitly during their time, and some of the things they say are open to interpretation.

28 Proponents of the hedonistic view of pleasure over pain include Bentham, *Introduction to the Principles of Morals and Legislation*; Mill, "Utilitarianism"; and Sidgwick, *Methods of*

I characterized AU in terms of *actual* well-being produced, one could also hold that it is *expected* well-being that matters (i.e., for each possible outcome, take the well-being that would be produced and multiply it by the probability of that outcome occurring, and then sum these). There are a number of other possible variations on the view, but the characterization above should do for my purposes.

To a moral theorist who takes a fundamental "first principle" approach to ethics, in which one searches for a highly simple, general principle, AU may be somewhat plausible. However, there are several obvious criticisms of the view. Arguably the most common one is that AU is easily "counterexampled" by finding intuitively immoral actions that AU prescribes. For instance, if an instance of murder (theft, torture, etc.) maximizes overall well-being—and it is not at all difficult to come up with cases like this—then AU says it is permissible. Other criticisms include that AU is too demanding in what it requires of you, too simplistic in its account of what is valuable, leaves no room for personal projects and relationships, and is not sensitive to how well-being is distributed. Proponents of AU have a number of responses to these worries, but this at least demonstrates a *prima facie* conflict between AU and our considered moral judgments.[29]

2.2. *Rossian Pluralism (aka "Commonsense Morality")*

RP, unlike AU, eschews the idea of a single fundamental principle underlying moral permissibility.[30] For this reason, it is difficult to give a concise statement of the view. However, the rough version is this: an action is morally permissible if and only if we do *not* have an *all-things-considered* duty not to do it, where this is determined by the balance of *pro tanto* duties, of which there is an irreducible plurality.[31] These duties can conflict and, for the resolution of such conflicts,

---

Ethics. For the preference-satisfaction account, see Smart, "An Outline of a System of Utilitarian Ethics"; and Harsanyi, "Rule Utilitarianism and Decision Theory." If Hare counts as a utilitarian, then he is also of the preference-satisfaction variety (*Moral Thinking*). This pluralistic account of well-being is often called the "objective list view." Proponents include Parfit, *Reasons and Persons*; and Hooker, *Ideal Code, Real World*.

29 For a critique of AU—in particular, for how badly it fits with our strongest moral convictions—see Williams, "A Critique of Utilitarianism"; Harsanyi, "Rule Utilitarianism and Decision Theory," 31; and Scheffler, *Consequentialism and Its Critics*, 1–13.

30 For the *locus classicus* of such a view, see Ross, *The Right and the Good*. Other proponents of RP include Nagel, "Fragmentation of Value"; and Audi, *The Good in the Right*.

31 Ross (*The Right and the Good*) unfortunately chose to label these as "*prima facie* duties," which suggests that they merely *seem* to be duties but may end up being morally irrelevant upon reflection. It is widely acknowledged that they are more accurately called "*pro tanto* duties," since they are considerations that count in favor of there being a duty proper but that might be outweighed by countervailing considerations. Kagan, *The Limits of Morality*, 17n13; Hooker, "Ross-Style Pluralism versus Rule-Consequentialism," 534n6.

are assigned specific weights (rather than being arranged in a lexical hierarchy). Furthermore, there is no fundamental principle governing the relative weights of the several principles to which we can appeal; rather, we must exercise opinion or judgment as to which *pro tanto* duties are defeated in which cases.[32] The *pro tanto* duties include, among other things, duties to keep one's promises, not to harm (steal, lie, etc.), and to promote well-being (giving priority to those to whom we bear special connections).

When compared to AU, the merits of RP should be obvious. Such a view, often described as the best regimentation of "commonsense morality," is very difficult, if not impossible, to "counterexample."[33] Anytime you construct a case that intuitively has some moral status that is not entailed by the current version of the theory, you can simply add a new *pro tanto* duty to your list that does cover it. Likewise, anytime several *pro tanto* duties conflict, this can be resolved by assigning the greatest weight to the one that intuitively should take precedence. However, there are some common criticisms of RP that are unrelated to how well it handles specific cases. The main charge, unsurprisingly, is that it is not systematic enough.[34] It paints morality as a "heap of unconnected duties," and its lack of an underlying principle governing which sort of actions are duties and their relative weights renders the view rather unexplanatory and unsatisfying from a theoretical perspective.[35] Likewise, the appeal to "judgment" for the purpose of resolving conflicts, rather than a principled explanation of the relative weights, is un-illuminating and perhaps even *ad hoc*. It is more like an evasion of the problem than a method.[36] If our intuitions are silent about the relative strengths of two competing *pro tanto* duties, there is simply no way to find out the truth of the matter.

2.3. *Rule Utilitarianism*

According to the version of RU that I favor, an action is morally permissible if and only if it is permitted by the rules whose general internalization has the great-

32 Ross, *The Right and the Good*, 19; Hooker, *Ideal Code, Real World*, 534.

33 Skelton, "William David Ross."

34 Joseph, *Some Problems in Ethics*; Rawls, *A Theory of Justice*; Skelton, "William David Ross."

35 McNaughton, "An Unconnected Heap of Duties?" 434. Ross, anticipating some of these "theoretical" worries, writes, "Loyalty to the facts is worth more than a symmetrical architectonic or a hastily reached simplicity" (*The Right and the Good*, 23). Of course, in order to avoid begging the question, he should replace the term "facts" with the term "evidence" or "considered moral judgments." However, if ETM is correct, then considerations of simplicity (and nonarbitrariness) must be included in our total evidence alongside our considered moral judgments.

36 Hare, *Moral Thinking*, 34.

est expected value in terms of overall well-being.[37] The notion of *internalization* here is meant to be distinct from that of full compliance; rather it is a matter of *accepting* the rules as the shared moral code in one's community, allowing for the possibility of failing to always live up to it. The expected value of the internalization of rules takes into account the costs of both implementation and maintenance.[38] There are more details that I am unable to currently fill in, so in this paper I will leave the account of RU as roughly stated above, while acknowledging that it is greatly in need of refinement.[39] Finally, RU also owes us some account of well-being, and the options are the same as they were for AU.

The attraction of RU—indeed its main motivation for those otherwise sympathetic toward utilitarianism or consequentialism in general—is that it can (allegedly) avoid common objections to AU while staying relatively systematic. I will make the case for this in more detail in section 3.2, albeit in terms of charity and eligibility. For now, I will mention the most common criticisms.

The greatest objection to RU, which is largely responsible for its unfavorable reputation among moral theorists in general, takes the form of a dilemma: either RU "collapses" into AU, in which case it loses its distinctiveness and gains all the problems associated with the latter, or else it is guilty of "incoherence," which could be understood as anything from logical inconsistency to being severely unmotivated as a version of consequentialism.[40] Roughly, the thought is that either the optimal set of rules prescribed by RU includes only the single rule of AU (i.e., "Choose the optimal action"), or else the rules tell you to take sub-

37 Proponents of RU include Harrod, "Utilitarianism Revised"; Harsanyi, "Rule Utilitarianism and Decision Theory"; and Brandt, *A Theory of the Good and the Right*. Harsanyi ("Rule Utilitarianism and Decision Theory," 32) credits the idea behind RU to Harrod ("Utilitarianism Revised") and the terms "act utilitarianism" and "rule utilitarianism" to Brandt (*Ethical Theory*, 380, 396). My characterization here is close to Hooker's formulation of his view, with some key differences (*Ideal Code, Real World*, 32). First and foremost, he is a rule *consequentialist* rather than a proponent of the narrower RU view. This is because he includes a principle of distribution, namely some priority for the worse off, rather than being merely aggregative, although he makes this choice rather tentatively (*Ideal Code, Real World*, 65). Also, he includes a clause favoring rules closer to conventional morality as a tiebreaker between otherwise optimal rules.

38 Hooker, *Ideal Code, Real World*.

39 I acknowledge the possibility that upon attempting to address such questions, RU may lose some of the eligibility that it appears to have. However, even once these details are filled in, it will still be far more eligible than RP.

40 On the idea that RU collapses into AU, see Lyons, who defends this thesis of the extension equivalence of AU and RU (*Forms and Limits of Utilitarianism*). On the idea that proponents of RU are guilty of "rule worship," a cardinal sin among consequentialists, see Smart, "An Outline of a System of Utilitarian Ethics," 10; and Kagan, *Normative Ethics*, 230.

optimal actions (even when you know they are suboptimal), which seems to abandon the spirit of consequentialism. Proponents of RU, or similar theories, have convincingly addressed how the first horn of the dilemma can be avoided, as suggested by the formulation of RU in terms of internalization rather than compliance—the general internalization of the single rule of AU simply would not have very good consequences.[41] The second horn, however, may still be worrisome, but it is beyond the scope of this paper to address it.[42]

3. THE STARTING POINT AND THE FIVE QUESTIONS

I will begin with the somewhat vague idea that we are morally required to "promote the good" and then will address several questions that arise when attempting to develop this idea with more clarity and precision. Readers who do not find such a starting point plausible are welcome to read the rest of this paper conditionally—that is, *if* you start out with this general idea about ethics, *then* considerations of charity and eligibility should move you toward RU. An alternative starting point for those whose moral sympathies are less consequentialist could be the vague idea that we are morally required to "respect persons," and I think a similar line of argument to the one in this paper could be made that leads from there to a view like contractualism.[43] Since other moral theories may be able to achieve a similarly good balance of charity and eligibility, proponents of such views still have much to gain from this discussion.

3.1. AU versus RP on the Five Questions

I will now introduce five questions that must be addressed in order to develop a moral theory from our starting point, the idea that we have a moral reason to promote the good. The first questions is:

Q1. Are there some things we may *not* do while promoting the good?

This is the question of whether there are, in the terminology of Kagan, moral *constraints*, or prohibitions on certain types of actions even when they have good consequences.[44] The second question is:

Q2. Are there some limits as to how much good we must promote?

41 Hooker, *Ideal Code, Real World*, 94.

42 For an answer to the charge of incoherence, see Hooker, *Ideal Code, Real World*, sec. 4.3.

43 For an overview of contractualism, see Ashford and Mulgan, "Contractualism." For a prominent contractualist view, see Scanlon, *What We Owe to Each Other*.

44 Kagan, *The Limits of Morality*, 4.

This is the question of whether there are, again in the terminology of Kagan, moral *options*, or limits on how demanding morality is, which give us room to pursue our own aims in life.[45] The third question is:

Q3. Is there anything worth promoting for its own sake other than well-being?

This is the question of the appropriate aims of morality. The fourth question is:

Q4. Can we favor those closest to us while promoting the good?

This is the question of whether morality permits some partiality toward those to whom we bear special relationships, or whether we must be completely impartial. Finally, the fifth question is:

Q5. While promoting the good, does it matter how well-being is distributed?

This question asks whether it is just the total amount of good promoted that matters, or whether it matters who gets it.

AU gives a negative answer to each of these five questions, which makes it highly counterintuitive. First, there are no moral constraints; anything that best promotes the good is permissible (and indeed required), no matter what intuitively immoral actions it involves (e.g., stealing, breaking promises, killing). Second, there are no moral options; we are required to *maximize* the good, which leaves us little to no room for any personal projects in life. No matter how generous we are with our time or money, we are still required to do more since there is always more good we can do. Third, there are no moral aims other than well-being. Everything else should only be pursued as a mere means to well-being, including intuitively valuable things such as knowledge, virtue, meaningful relationships, and personal accomplishments. Fourth, we are not allowed any partiality toward those closest to us. We must be completely impartial in our good-promoting actions, giving absolutely equal consideration to the well-being of all, strangers and loved ones alike. Finally, the distribution of well-being does not matter; all that matters is that we produce as much well-being as possible, no matter how unequally it is distributed or whether it makes its way to those deserving or undeserving. Given all these implications, AU conflicts greatly with some of our strongest moral convictions, making it an extremely uncharitable theory.

RP, on the other hand, gives an affirmative answer to each of these questions, which makes it highly intuitive. First, there are moral constraints, since there

45 Kagan, *The Limits of Morality*, 3.

are *pro tanto* duties other than our duty to promote the good. We must, in ordinary circumstances, keep our promises, not steal, not harm, etc., and only when there are strong enough countervailing considerations are we permitted to violate these constraints. Second, there are moral options, since the *pro tanto* duty to promote the good, according to RP, is plausibly interpreted in terms of *satisficing*, or doing "enough" good, rather than maximizing.[46] As long as you have done enough to satisfy the duty of beneficence, you have freedom to pursue your own projects. Third, there are aims to promote other than well-being, since in addition to countenancing a plurality of duties, RP is also pluralistic about the good. We should aim to promote well-being, but we should also aim to promote knowledge, virtue, and various other intuitively valuable goods. Fourth, we are allowed some partiality toward those to whom we bear special relationships. We can favor the well-being of ourselves and our loved ones, provided that we not disregard others' interests entirely. Finally, the distribution of well-being matters. Two plausible candidates for distribution principles include one according to desert—for instance, the "allocation of pleasure to the virtuous"—and priority toward the worse off.[47] Given these implications, RP fits very well with our moral convictions and is hence an extremely charitable theory.

Unfortunately, the high degree of charity that RP achieves comes at a very steep price in terms of eligibility. To demonstrate this, I will discuss how each affirmative answer RP gives affects the definition of *permissible*$_{RP}$. First, we begin with the following simple definition:

[RP1] *being an action that promotes the good*

Now, adding in moral constraints, we get:

[RP2] *being an action that promotes the good without* C_1 *or* C_2 *or … or* C_n

where each C_i corresponds to a constraint-violating act-type (e.g., perhaps C_2 is *breaking a promise*). However, recall that RP's constraints are non-absolute; most constraints have a set of exception clauses for when they are outweighed by stronger moral considerations. Thus, we must add a layer of complication to the definition:

[RP3] *being an action that promotes the good without* $[C_1$ *and not* $(E_{1,1}$ *or* $E_{1,2}$ *or … or* $E_{1,h})]$ *or* $[C_2$ *and not* $(E_{2,1}$ *or* $E_{2,2}$ *or … or* $E_{2,k})]$ *or … or* $[C_n$ *and not* $(E_{n,1}$ *or* $E_{n,2}$ *or … or* $E_{n,m})]$

where each $E_{i,j}$ is the *jth* exception to the *ith* constraint. For instance, if C_2 is

46 On satisficing, see Slote and Pettit, "Satisficing Consequentialism."

47 Ross, *The Right and the Good*, 140.

breaking a promise, $E_{2,3}$ might be *saving someone from severe distress*. Next, we must account for moral options, by interpreting "promotes" as *satisfices:*

[RP4] *being an action that leads to at least quantity Q of good without* $[C_1$ *and not* $(E_{1,1}$ *or* $E_{1,2}$ *or … or* $E_{1,h})]$ *or* $[C_2$ *and not* $(E_{2,1}$ *or* $E_{2,2}$ *or … or* $E_{2,k})]$ *or … or* $[C_n$ *and not* $(E_{n,1}$ *or* $E_{n,2}$ *or … or* $E_{n,m})]$

where *Q* is some complete specification of what counts as "enough" good. Next, since RP is pluralistic about the good, we must fully specify the several components of the good and assign relative weights to each to account for trade-offs (e.g., to determine how much knowledge is worth promoting over a certain amount of well-being):

[RP5] *being an action that leads to at least quantity Q of the sum:* $(W_1G_1 + W_2G_2 + \ldots + W_vG_v)$ *without* $[C_1$ *and not* $(E_{1,1}$ *or* $E_{1,2}$ *or … or* $E_{1,h})]$ *or* $[C_2$ *and not* $(E_{2,1}$ *or* $E_{2,2}$ *or … or* $E_{2,k})]$*or … or* $[C_n$ *and not* $(E_{n,1}$ *or* $E_{n,2}$ *or … or* $E_{n,m})]$

where each G_i is a quantity of one of the goods worth promoting for its own sake and each W_i is its appropriate weight. Next, we must specify the details of its partiality:

[RP6] *being an action that leads to at least quantity Q of the sum:* $(W_1G_1 + W_2G_2 + \ldots + W_vG_v)$*, giving those in group* S_1 *priority* P_1*, those in group* S_2 *priority* P_2*, … , and those in group* S_w *priority* P_w*, without* $[C_1$ *and not* $(E_{1,1}$ *or* $E_{1,2}$ *or … or* $E_{1,h})]$ *or* $[C_2$ *and not* $(E_{2,1}$ *or* $E_{2,2}$ *or … or* $E_{2,k})]$ *or … or* $[C_n$ *and not* $(E_{n,1}$ *or* $E_{n,2}$ *or … or* $E_{n,m})]$

where each S_i is a set of individuals related to the agent in a morally relevant way and each P_i is the appropriate weight of prioritizing the well-being of those in S_i. For instance, perhaps S_1 is the agent's singleton and P_1 is the greatest prioritizing weight, S_2 includes the agent's closest friends and family members and P_2 is the second greatest weight, etc. Finally, we must account for RP's distribution principles, leading to our completed definition:

[RP] *being an action that* $[$*leads to at least quantity Q of the sum:* $(W_1G_1 + W_2G_2 + \ldots + W_vG_v)$*, where the well-being of those in group* S_1 *is given priority* P_1*, … , and the well-being of those in group* S_w *is given priority* P_w*, giving weight* D_1 *to the deserving,* D_2 *to those worse off, … , and* D_r *to those with feature* $F_r]$ *without* $[C_1$ *and not* $(E_{1,1}$ *or* $E_{1,2}$ *or … or* $E_{1,h})]$ *or* $[C_2$ *and not* $(E_{2,1}$ *or* $E_{2,2}$ *or … or* $E_{2,k})]$*or … or* $[C_n$ *and not* $(E_{n,1}$ *or* $E_{n,2}$ *or … or* $E_{n,m})]$

where each D_i is the appropriate weight of distributing well-being in favor of those with feature F_i.

Now that we have gone through how RP's answers to the five questions affect the definition of *permissible*$_{RP}$, we can compare it to the definition of *permissible*$_{AU}$, which is something like the following:

> [AU] *being an action that leads to as much overall well-being for all, equally considered, as possible*

Clearly [RP] is far longer and vastly more complex than [AU]. Even though [AU] references complex phenomena such as well-being, whose canonical definition may itself be relatively complex, [RP] does so as well, while also referencing a lot more. Furthermore, [RP] contains a significant amount of arbitrariness, since it assigns many arbitrary weights and includes many arbitrary exception clauses. For instance, why think that we are only required to promote quantity Q of good, rather than $Q + k$ or $Q - k$? Also, why think we may violate constraint C_2 only to prevent at least quantity Q' of bad, rather than $Q' + k$ or $Q' - k$? Countless properties are extremely similar to the one expressed by [RP] that differ only in assigning a slightly different value to one of these variables, or in swapping in or out some one or two exception clauses. Any choice between them will be completely arbitrary. The fact that [RP] is extremely complex and arbitrary suggests that *permissible*$_{RP}$ is extremely unnatural and much less natural than *permissible*$_{AU}$.

Let me briefly summarize. The theories AU and RP give opposite answers to each of the five questions, where one answer leads to a much simpler and less arbitrary but highly counterintuitive theory, while the other answer leads to a highly intuitive but extremely complex and arbitrary theory. Whether I got the details of *permissible*$_{RP}$ exactly correct is debatable, but that such a property will inevitably be extremely complex and arbitrary is not. Thus, if we begin at our starting point with a vague requirement to promote the good and address the five questions, considerations of charity and eligibility seem to pull strongly in opposite directions.

3.2. *How RU Answers the Five Questions*

I will now show how RU achieves a moderately high degree of both charity and eligibility. Before we see how RU fares in addressing the five questions, let us remind ourselves how it scores on eligibility. The definition of *permissible*$_{RU}$ looks something like:

> [RU] *being an action that is permitted by the rules whose general internal-*

ization has the greatest expected value in terms of overall well-being for all, equally considered

As I admitted in section 2.3, a lot of details still need to be filled in, and doing so may lead to greater complexity and arbitrariness than is now apparent. However, I think it is clear that while [RU] is not as simple or nonarbitrary as [AU], it is far simpler and less arbitrary than [RP] and will continue to be even once these details are filled in. Thus, *permissible*$_{RU}$ is much more natural than *permissible*$_{RP}$.

However, unlike AU, RU can maintain this relatively high degree of eligibility without sacrificing too much charity. This is because RU gives the intuitive answers to each of the five questions, the same affirmative answers that RP gives. First, RU will include moral constraints, since the rules whose internalization has the greatest expected utility (henceforth, the "optimal rules") will not simply consist of one rule that says "Maximize well-being," but will instead consist of a plurality of (plausibly non-absolute) rules such as "Keep your promises in ordinary circumstances" and "Don't harm an innocent person (except to prevent a disastrous outcome)." The general internalization of constraints against dishonesty, promise-breaking, and violating property rights in ordinary circumstances is necessary to secure trust and other beneficial expectation effects.[48] Likewise, constraints against harm help to avoid miscalculation and abuse, and constraints against free-riding produce beneficial coordination effects.[49] In general, there is great social utility in the general acceptance of constraints and their correlative rights.[50]

Second, RU will plausibly be much less demanding than AU, with its optimal rules leaving people with options to pursue their own aims and projects.[51] Whatever rule in the optimal rule set is associated with promoting well-being will plausibly be stated in terms of satisficing rather than maximizing. This is because the costs of getting a hyper-demanding rule (e.g., "Be altruistic to the point of diminishing marginal utility") internalized among the general population and maintaining it would be extremely high.[52] Even if people could be

48 Harsanyi, "Rule Utilitarianism and Decision Theory," 32–33; Hooker, *Ideal Code, Real World*, 77.

49 Kagan, *The Limits of Morality*, 33–34.

50 For an extensive discussion of how RU and similar indirect consequentialist views justify constraints, see Hooker, *Ideal Code, Real World*, ch. 6. For a thorough discussion of whether such constraints are plausibly absolute or non-absolute, also see Hooker, *Ideal Code, Real World*, sec. 6.4.

51 See Hooker, *Ideal Code, Real World*, ch. 8. For dissent regarding RU's ability to avoid excessive demandingness, see Kagan, *The Limits of Morality*, 35.

52 Hooker, *Ideal Code, Real World*, 78–79.

convinced of the moral authority of such a rule, which is dubious, they would constantly fail to live up to it and thereby alienate themselves from morality in general and perhaps other important moral rules in particular.[53] Ironically, it is plausible that people would end up being motivated to do *more* good if there's a less demanding rule concerning how much good they are required to do that leaves room for supererogatory action beyond that.[54] After all, sometimes you get more with honey than with vinegar.

Third, RU's rules governing good-promotion will not tell you to just aim at promoting well-being for its own sake but other things as well, including knowledge, virtue, and justice. This is an extension of the lesson drawn from the so-called paradox of hedonism, which is the observation that "adopting as one's exclusive ultimate end in life the pursuit of maximum happiness may well prevent one from having certain experiences or engaging in certain sorts of relationships or commitments that are among the greatest sources of happiness."[55] This sort of "paradox" can be generalized into what we might call the "paradox of welfarism"—in other words, adopting overall well-being as the only direct aim in our everyday lives will likely result in less overall well-being. This is because such an attitude would preclude us from aiming *directly* at things like accomplishments, scientific or philosophical discovery, meaningful relationships, and self-improvement; these other things would be treated as purely instrumental, worth pursuing only if our direct utility calculations yield the right verdict. Such a single-minded way of deliberating, apart from being wildly impractical, seems much less conducive to overall well-being than the alternative, namely pursuing a reasonable plurality of aims.

Fourth, RU would plausibly permit some degree of partiality, since internalizing practical rules that allow or even mandate some degree of partiality would have better consequences in terms of overall well-being. Given human psychology, there would be significant costs in attempting to get and keep fully impartial practical rules internalized.[56] Furthermore, there are certain benefits that can best be secured through partiality, including personal accomplishments, which require favoring your own interests, and meaningful relationships, which require favoring the interests of those close to you.[57] In general, overall well-being is better promoted when we follow rules that prescribe some degree of partiali-

53 Kagan, *The Limits of Morality*, 35.

54 Kagan, *Normative Ethics*, 225.

55 Railton, "Alienation, Consequentialism, and the Demands of Morality," 140.

56 Hooker, *Ideal Code, Real World*, 140.

57 Hooker, *Ideal Code, Real World*, 139.

ty toward ourselves and those closest to us, since those are the interests we are most familiar with and in the best position to affect.[58]

Finally, RU's rules governing good-promotion will plausibly take distribution into account. As Hooker explains, "utilitarians have to trade off the diminishing marginal utility of material goods against the need for economic incentives."[59] The former consideration calls for a distribution principle that gives priority to the worse off, while the latter calls for a distribution principle based on desert.[60] Thus the optimal rule set will plausibly require us to give priority to the worse off and the virtuous when promoting well-being, as such a practice is much more conducive to overall well-being than the alternative.

Let us take stock of how RU fares when addressing the five questions. Given that the affirmative answers RU gives are not *assumed* as part of the theory, like they are on RP, but are instead *derived* from the theory together with empirical considerations, RU is able to secure a degree of eligibility that far surpasses that of RP. It avoids numerous complications and countless arbitrary choices about exactly how to assign specific values and where to draw certain lines (between, for instance, cases that are exceptions to a certain constraint and those that are not). However, the (complex and arbitrary) contingent, empirical facts being what they are, the theory yields a highly complex set of practical rules that map onto our considered moral judgments in a fairly comfortable manner. Again, this fit will be far from perfect—for instance, RU may still be a bit more demanding than we expected morality to be—but there is a world of difference between the charity of RU and that of AU. Thus, if we begin our moral theorizing from the starting point of a vague requirement to promote the good and address the five questions in order to clarify and precisify this intuition, RU looks like a very promising moral theory, securing a nice balance of charity and eligibility.

58 Jackson makes this point in a particularly compelling way using his "crowd control" thought experiment ("Decision-Theoretic Consequentialism and the Nearest and Dearest Objection," 474).

59 Hooker, *Ideal Code, Real World*, 64.

60 An additional merit of this sort of approach is that it is consistent with a thoroughgoing skepticism about basic, non-consequentialist desert. See Caruso ("Skepticism about Moral Responsibility") for discussion. The sort of desert invoked at the practical level is merely consequentialist. The virtuous do not deserve a benefit in any deep sense that requires a controversial sort of moral responsibility or free will. Rather, they "deserve" it because treating them so is part of a useful practice, providing incentive effects that are conducive to overall well-being.

4. OBJECTIONS

In this section, I will consider and address three objections. They all concern whether RU is really as charitable as I suggest. Some of my responses will be rather concessive, granting in many cases that RU's fit with our considered judgments is far from perfect. However, when we keep ETM and considerations of eligibility in mind, I think RU's mismatch with usage is far from decisive.

4.1. The "Wrong" Rules and Lines in All the "Wrong" Places

First, one might grant that I have successfully shown that RU can make room for *certain* constraints, options, pluralistic aims, degrees of partiality, and principles of distribution, but not that I have shown that RU can get the *intuitively correct* ones. Given the contingent, empirical facts, the optimal constraints, weights, etc. might be somewhat different from what we intuitively think. For instance, RU's line that marks where satisficing ends, where the constraint against promise-breaking gives way to the exception for preventing distress, or where some quantity of virtue outweighs some quantity of well-being, might not be exactly where our moral intuitions want it to be. Thus RU may end up being *much* less charitable than RP, since the latter can take the constraints, weights, lines, etc. to be *exactly* as they intuitively seem (except when our intuitions are inconsistent), while the practical rules of RU are hostage to contingent, empirical facts.

In response to this objection, I will first point out that even getting *some* constraints, options, etc. is still a considerable achievement and puts RU light-years ahead of AU with respect to charity. The counterintuitiveness of AU is altogether a difference in kind, given that it has *no* constraints, options, etc., whereas the counterintuitiveness of RU's imperfect (with respect to our considered judgments) constraints and line placements is just a matter of degree. That there are *some* constraints, options, etc. may be close to a "Moorean fact" about moral permissibility, whereas we seem to be more open to revising exactly where we draw certain lines.

Second—and this will be somewhat of a recurring theme in my responses to objections—I expect the superior eligibility of RU (over RP) to pick up the slack wherever its charity falls short. After all, ETM, in contrast to a charity-only metasemantics, gives revisionary theories a chance of being true despite some mismatch with usage. I see RU as an instance of this general idea; the superior naturalness of $permissible_{RU}$ makes up for its less than perfect fit with our considered judgments (e.g., some lines in the intuitively wrong places).

4.2. The Contingency of the Rules and Otherworldly "Counterexamples"

Next, one might grant that RU can yield the intuitively correct (or close enough) practical rules in the *actual* world, given the *actual* facts about human psychology and our environment. However, when we consider other possible worlds with other such possible facts about agents and environments, the practical rules derived from RU may be drastically different and highly counterintuitive.[61] For instance, perhaps we can imagine alternative agent psychologies or laws of nature such that, in those worlds, internalizing rules that permit or mandate torture, theft, etc. has a very high expected value in terms of overall well-being. Thus RU, though it can avoid (most of) the easy, *this*-worldly counterexamples to AU, may still be subject to damning, *other*worldly counterexamples. After all, the constraints, weights, lines, etc. on RU, unlike on RP, are an entirely *contingent* matter.

In response, I think that not all counterexamples are created equal in terms of their theoretical import. When it comes to the metasemantic constraint of charity, fitting with our usage of the term over typical and familiar cases counts more in reference determination than fitting with our usage of the term over far-fetched and unfamiliar cases. Thus if there is a candidate referent that fits somewhat poorly with our dispositions to apply the term to extremely atypical cases, but otherwise fits very well, then considerations of charity should not disqualify it from being the referent of that term. Hence, the fact that RU can handle (most of) our considered judgments about typical, actual, and *nearby* possible cases gives it the degree of charity it needs to be a strong competitor in the battle of theory choice. Considerations of eligibility can take it the rest of the way.

4.3. The Right Rules but for All the "Wrong" Reasons

Finally, one may grant that RU does a good enough job at fitting our considered judgments about the practical rules of morality but then complain that it does a bad job at fitting our judgments about *why* those are the correct rules. Perhaps RU can correctly account for the fact that we are required to keep our promises, promote virtue, pursue meaningful relationships, etc., but its explanation for these facts may be highly counterintuitive. We typically think that promise-keeping, virtue, meaningful relationships, etc. are valuable *in themselves*, whereas RU holds that their value is derivative, wholly explained by their (indirect) relationship to the value of well-being. In general, if we have a constraint against performing

61 For a recipe for constructing some such (modally distant) counterexamples to RU or rule consequentialism in general, see Podgorski, "Wouldn't It Be Nice?" Thanks to an anonymous referee for this reference.

actions of a certain type, that's because it seems like there is something wrong with those actions *in themselves* rather than merely because they are prohibited by the optimal rules. Thus even if RU can yield the intuitively correct practical rules, it does so for the intuitively wrong reasons. The revisionary nature of RU's moral explanations means that the theory must take a hit with respect to charity.

I will give two responses to this objection, the first combative and the second concessive. First, even when we think about our judgments of moral explanations, it is not obvious that a view like RP is that much more charitable than RU. While RP, unlike RU, can agree with our intuitions that constraints, pluralistic aims, etc. have fundamental moral value, we also seem to have the intuition that there is some deeper, unifying explanation behind ordinary moral rules. This is evident from the fact that, long before the notions of naturalness or eligibility were anywhere on the scene, moral philosophers were complaining that views like RP are too unsystematic—mere "shopping lists" of disconnected principles and unexplained arbitrary weights. These two intuitions, that multiple sorts of things are fundamentally valuable and that there is a unifying explanation behind all of morality, seem to be in conflict. RP does justice to the former and RU the latter. Thus, even taking into account our convictions about moral explanations, there may not be as big a difference in charity between RP and RU as the objection suggests.

Second, even if the objection is correct that RU's moral explanations are somewhat revisionary, this does not mean we should not accept them. If we are realists about ethics, just like realists about anything, we should be open to somewhat surprising explanations behind ordinary phenomena. We have learned from modern science that there are all sorts of extremely surprising and counterintuitive explanations (e.g., atomic theory, quantum mechanics) behind the behavior of ordinary things at the macroscopic level. RU can be seen as just another instance of this general theme, albeit in the moral domain. Once again, charity is not the be-all and end-all; eligibility must be given its due weight.

5. CONCLUSION

Before I close, it is worth briefly discussing the costs of denying ETM in metaethics, or of downplaying the strength of the eligibility constraint to the point where a view like RP could end up achieving the best balance of charity and eligibility in spite of its low eligibility. If charity were given near full authority in the metasemantics of "morally permissible," then if there were to be two or more equally charitable interpretations, the term would be semantically indeterminate between them. If our moral intuitions, together with those of the rest

of our linguistic community, were split or undecided on some matter—for instance, on the presence or absence of some particular constraint or the value of some particular weight—then there would be no fact of the matter as to what is morally permissible in cases that turn on this difference. Furthermore, if we were to encounter a moral theorist from an alternative linguistic community—or whose position could best be understood by reference to a corresponding hypothetical linguistic community—with different moral convictions about certain constraints or weights, then many of our disputes with her about what is morally permissible would be verbal.[62] Thus there are significant metaphilosophical costs of downplaying the role of eligibility.

However, it should be noted that even if we do adopt ETM there is no guarantee of securing shared reference for every linguistic community with a term that plays the permissibility role. For instance, a community of committed act utilitarians, whose usage of "morally permissible" aligns very closely with what AU entails, may refer to $permissible_{AU}$ after all, since it is this property that will best balance charity and eligibility in that community. This is, however, the exception that proves the rule. It is only because the usage of these act utilitarians is so vastly different from our own that we end up expressing distinct properties by our respective moral terms.[63] For any linguistic community whose usage is in the vicinity of what we consider "commonsense morality," shared reference will be secured to $permissible_{RU}$ due to its decent fit and high degree of eligibility. Thus most moral disputes will still come out as nonverbal.

My concession that ETM, under the assumptions about naturalness N1–N4 outlined in section 1, does not provide the strong *guarantee* of shared reference for all possible moral communities may seem to undermine the main motivation for ETM in metaethics, namely its use as a general solution to the Moral Twin Earth challenge. If this concession is too much for some theorists, then this

62 This is assuming Hirsch's ("Physical-Object Ontology, Verbal Disputes, and Common Sense") account of verbal disputes, which is motivated by its ability to remain faithful to Burge's ("Individualism and the Mental") social externalist insight, namely that what we mean is partly determined by the patterns of usage in our wider linguistic community (i.e., meaning is not a completely private matter). A dispute's being verbal on this account does not require that the disputants mean different things (since members of the same linguistic community typically speak a shared language) but only that the hypothetical linguistic communities with the parties' differing usages would mean different things.

63 Perhaps it could be argued that $permissible_{AU}$ is not in fact a candidate referent for "permissible," since it cannot play all of the permissibility role in our thought and discourse, which includes action guidance. If so, then perhaps $permissible_{RU}$ is the most natural candidate after all, in which case there may be more of a guarantee of shared reference due to reference magnetism, at least to the extent that RU achieves a high enough degree of charity for every possible community with a term that plays the permissibility role.

may give them more reason to reject one or more of N1–N4. However, I think securing a reasonable amount of shared reference—in particular, for all moral communities in the vicinity of "commonsense morality"—is motivation enough.

In this paper, I have argued that RU is a very promising theory if we adopt a metasemantics that includes ETM. On RU, the moral property comes out as fairly simple and nonarbitrary, especially when compared to views like RP. Since her moral property is relatively natural, the proponent of RU can reap the benefits of reference magnetism, which includes limiting semantic indeterminacy and securing shared reference between alternative linguistic communities with somewhat diverging usages, thus avoiding verbal disputes. Unlike its rival AU, RU secures this high degree of eligibility without sacrificing too much by way of charity. Hence, RU should be taken very seriously by any moral philosopher who aims to "carve nature at its joints."[64]

University of California, Santa Barbara
dmokriski@gmail.com

REFERENCES

Ashford, Elizabeth, and Tim Mulgan. "Contractualism." *Stanford Encyclopedia of Philosophy* (Summer 2018). https://plato.stanford.edu/archives/sum2018/entries/contractualism/.

Audi, Robert. *The Good in the Right: A Theory of Intuition and Intrinsic Value*. Princeton: Princeton University Press, 2004.

Bentham, Jeremy. *Introduction to the Principles of Morals and Legislation*. London: T. Payne and Son, 1789.

Boyd, Richard. "How to Be a Moral Realist." In *Essays on Moral Realism*, edited by Geoffrey Sayre-McCord, 181–228. Ithaca, NY: Cornell University Press, 1988.

Brandt, Richard B. *Ethical Theory*. Upper Saddle River, NJ: Prentice-Hall, 1959.

———. *A Theory of the Good and the Right*. Buffalo, NY: Prometheus Books, 1979.

Brink, David O. *Moral Realism and the Foundation of Ethics*. Cambridge: Cambridge University Press, 1989.

64 Thanks to Arnel Blake Batoon, Sherri Lynn Conklin, Ryan Jenkins, Daniel Story, and the attendees of my talks at UC Santa Barbara's Graduate Colloquia and Cal Poly's Philosophy Research Workshop for useful feedback. And very special thanks to Dan Korman as well as two anonymous reviewers for extremely helpful comments on earlier drafts.

Burge, Tyler. "Individualism and the Mental." *Midwest Studies in Philosophy* 4, no. 1 (September 1979): 73–122.

Caruso, Gregg. "Skepticism about Moral Responsibility." *Stanford Encyclopedia of Philosophy* (Spring 2018). https://plato.stanford.edu/archives/spr2018/entries/skepticism-moral-responsibility/.

Dorr, Cian, and John Hawthorne. "Naturalness." In *Oxford Studies in Metaphysics*, vol. 8, edited by Karen Bennett and Dean W. Zimmerman, 1–70. Oxford: Oxford University Press, 2013.

Dunaway, Billy, and Tristram McPherson. "Reference Magnetism as a Solution to the Moral Twin Earth Problem." *Ergo* 3, no. 25 (2016): 639–79.

Edwards, Douglas. "The Eligibility of Ethical Naturalism." *Pacific Philosophical Quarterly* 94, no. 1 (March 2013): 1–18.

———. "Naturalness, Representation, and the Metaphysics of Truth." *European Journal of Philosophy* 21, no. 3 (September 2013): 384–401.

Gettier, Edmund. "Is Justified True Belief Knowledge?" *Analysis* 23, no. 6 (June 1963): 121–23.

Goodman, Nelson. *Fact, Fiction, and Forecast.* Cambridge, MA: Harvard University Press, 1955.

Guigon, Ghislain. "Overall Similarity, Natural Properties, and Paraphrases." *Philosophical Studies* 167, no. 2 (January 2014): 387–99.

Hare, R. M. *Moral Thinking: Its Levels, Method, and Point*. Oxford: Oxford University Press, 1981.

Harrod, R. F. "Utilitarianism Revised." *Mind* 45, no. 178 (April 1936): 137–56.

Harsanyi, John C. "Rule Utilitarianism and Decision Theory." *Erkenntnis* 11, no. 1 (January 1977): 25–53.

Hawthorne, John. "Craziness and Metasemantics." *Philosophical Review* 116, no. 3 (July 2007): 427–40.

Hirsch, Eli. *Dividing Reality.* Oxford: Oxford University Press, 1993.

———. "Physical-Object Ontology, Verbal Disputes, and Common Sense." *Philosophy and Phenomenological Research* 70, no. 1 (January 2005): 67–97.

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

———. "Ross-Style Pluralism versus Rule-Consequentialism." *Mind* 105 no. 420 (October 1996): 531–52.

Horgan, Terence, and Mark Timmons. "New Wave Moral Realism Meets Moral Twin Earth." *Journal of Philosophical Research* 16 (1991): 447–65.

Jackson, Frank. "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101, no. 3 (April 1991): 461–82.

Joseph, H. W. B. *Some Problems in Ethics*. Oxford: Clarendon Press, 1931.

Kagan, Shelly. *The Limits of Morality*. Oxford: Oxford University Press, 1989.
———. *Normative Ethics*. Abingdon, UK: Routledge, 1998.
Kripke, Saul. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press, 1982.
Lewis, David. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61, no. 4 (December 1983): 343–77.
———. "Putnam's Paradox." *Australasian Journal of Philosophy* 62, no. 3 (September 1984): 221–36.
Lyons, David. *Forms and Limits of Utilitarianism*. Oxford: Clarendon Press, 1965.
McNaughton, David. "An Unconnected Heap of Duties?" *Philosophical Quarterly* 46, no. 185 (October 1996): 433–47.
Mill, John Stuart. "Utilitarianism." In *Oxford Philosophical Texts*, edited by Roger Crisp. Oxford: Oxford University Press, 1998.
Nagel, Thomas. "Fragmentation of Value." In *Mortal Questions*, 128–41. Cambridge: Cambridge University Press, 1979.
Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.
Podgorski, Abelard. "Wouldn't It Be Nice? Moral Rules and Distant Worlds." *Noûs* 52, no. 2 (June 2018): 279–94.
Putnam, Hilary. "Realism and Reason." *Proceedings and Addresses of the American Philosophical Association* 50, no. 6 (August 1977): 483–98.
Railton, Peter. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs* 13, no. 2 (Spring 1984): 134–71.
———. "Moral Realism." *Philosophical Review* 95, no. 2 (April 1986): 163–207.
Rawls, John. *A Theory of Justice*. Cambridge, MA: Harvard University Press, 1971.
Ross, W. D. *The Right and the Good*. Oxford: Clarendon Press, 1930.
Scanlon, T. M. *What We Owe to Each Other*. Cambridge, MA: Belknap Press of Harvard University Press, 1998.
Schaffer, Jonathan. "Two Conceptions of Sparse Properties." *Pacific Philosophical Quarterly* 85, no. 1 (March 2004): 92–102.
Scheffler, Samuel, ed. *Consequentialism and Its Critics*. Oxford: Oxford University Press, 1988.
Sider, Theodore. "Criteria of Personal Identity and the Limits of Conceptual Analysis." *Philosophical Perspectives* 15 (2001): 189–209.
———. "Ontological Realism." In *Metametaphysics: New Essays on the Foundations of Ontology*, edited by David Chalmers, David Manley, and Ryan Wasserman, 384–423. Oxford: Clarendon Press, 2009.
———. *Writing the Book of the World*. Oxford: Clarendon Press, 2011.
Sidgwick, Henry. *Methods of Ethics*, 7th ed. London: Macmillan, 1907.

Singer, Peter. "Is Act-Utilitarianism Self-Defeating?" *Philosophical Review* 81, no. 1 (January 1972): 94–104.

Skelton, Anthony. "William David Ross." *Stanford Encyclopedia of Philosophy* (Summer 2012). https://plato.stanford.edu/archives/sum2012/entries/william-david-ross/.

Slote, Michael, and Philip Pettit. "Satisficing Consequentialism." *Proceedings of the Aristotelian Society* 58 (1984): 139–76.

Smart, J. J. C. "An Outline of a System of Utilitarian Ethics." In *Utilitarianism: For and Against*, by J. J. C. Smart and Bernard Williams, 3–76. Cambridge: Cambridge University Press, 1973.

Van Roojen, Mark. "Knowing Enough to Disagree." In *Oxford Studies in Metaethics*, vol. 1, edited by Russ Shafer-Landau, 161–94. Oxford: Oxford University Press, 2006.

Weatherson, Brian. "The Role of Naturalness in Lewis's Theory of Meaning." *Journal of the History of Analytic Philosophy* 1, no. 10 (2013): 1–19.

———. "What Good Are Counterexamples?" *Philosophical Studies* 115, no. 1 (July 2003): 1–31.

Williams, Bernard. "A Critique of Utilitarianism." In *Utilitarianism: For and Against*, by J. J. C. Smart and Bernard Williams, 77–150. Cambridge: Cambridge University Press, 1973.

Williams, J. Robert G. "Eligibility and Inscrutability." *Philosophical Review* 116, no. 3 (July 2007): 361–99.