Meaning and Morality

by

Brian Craig Robinson

A dissertation submitted to the Graduate Faculty in Philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2011

This manuscript has been read and accepted for the Graduate Faculty in Philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Date	Gary Ostertag Chair of Examining Committee
Date	Iakovos Vasiliou Executive Officer, Philosophy

Stephen Neale, Advisor

Michael Devitt

Jesse Prinz

Rohit Parikh

THE CITY UNIVERSITY OF NEW YORK

Abstract

Meaning and Morality

by

Brian Robinson

Advisor: Professor Stephen Neale

Our ability to use language meaningfully derives in large part from our status as moral agents. The role of value and normativity cannot be separated from meaning and language use.

Paul Grice's seminal work on implicature introduces the intuitive distinction between what is said and what is meant. What is implicated is supposed to fill this gap. As has been previously noted, Grice's theory relies heavily on unexplicated notions of rationality, cooperation, and intentions. This dissertation seeks to explore these notions, grounding them in a theory of moral psychology, and then examining what modifications to the theory of implicature are then needed.

I begin with a review of suggestive, but unsystematic remarks left by Grice on psychology and ethics. From these comments, I construct a novel, quasi-Gricean theory of moral psychology I call Rational Virtue Theory, which is an egoistic moral psychology that falls under the rubric of virtue ethics, given its emphasis on *eudaimonia* (or happiness). This theory posits various End Selection Rules and Behavioral Principles. End Selection Rules are designed to guide one in constructing one's goals for life in pursuit of *eudaimonia*. Behavioral Principles are *ceteris paribus* strategies for action capable of repetition and replication, and which are stable and rational.

With this theory of moral psychology, I take on the issue of cooperation, first arguing for its status as a behavioral principle. Cooperation in language use, however, I contend is distinct from cooperation in general. I then argue that there are two distinct notions of cooperation in language use: conversational cooperation and communicative cooperation. Though Grice appears to endorse the former, I assert that the latter is the actual source of Grice's conversational maxims and necessary for implicature and successful communication. Next, I argue for a variety of modifications to the theory of implicature, including allowing for moral implicatures and the role of moral maxims in working out what a speaker has implicated. As it turns out, the gap between what is said and what is meant is much larger than Grice initially conceived of, and understanding the role of rationality and normativity in language use helps fill in more of that gap.

Acknowledgments

There were many people who were instrumental in helping this dissertation reach completion. First and foremost, I wish to thank my dear wife, Lesley, for all her love, support, and endless patience. I couldn't have done this without her.

My advisor, Stephen Neale, was a source of incredible aid and insight throughout this process. Thank you for all the feedback and stimulating conversations. You also planted the seed of the idea that grew into this dissertation.

I also wish to thank all my committee members – Jesse Prinz, Michael Devitt, Rohit Parikh, and Gary Ostertag for their comments and engaging questions.

Finally, thank you to my many graduate student colleagues who read drafts or gave feedback. In particular I want to thank my dissertation group – Mark Alfano, Dan Shargel, Myrto Mylopoulos, and Todd Beatie – for meeting regularly and reading so much.

Table of Contents

Chapter 1: Introduction

The central claim that I will advocate is that we simply cannot separate value and normativity from meaning and language use. Our ability to use language meaningfully derives in large part from our status as moral agents.

Paul Grice's theory of implicature (1975a, 1986, 1989, 1998) is well known and widely discussed among linguists, philosophers of language, and cognitive scientists. By comparison, his work on ethics and philosophical psychology is virtually unknown. This is unfortunate, not just because of the latter's intrinsic philosophical interest, but because Grice himself saw it as providing foundations upon which his work on meaning and implicature rested. The full story is not spelled out anywhere in Grice's published work, and a considerable amount of detective work is needed to piece together from Grice's hints anything that would come close to constituting a theory

(particularly by the exacting standards Grice seems to have held his own work to). But my aim here is not to reconstruct anything I claim is Grice's own theory.

For the first major aim of this project, I will develop a new, egoistic theory of moral psychology that is a type of virtue ethics. This theory, which I call Rational Virtue Theory, will be Gricean in the sense it will be based upon suggestions left by Grice. However, it will not be Gricean in a different sense. I am not attempting to reconstruct Grice's own, unpublished theory of moral psychology based on his existing writings. I suspect that Rational Virtue Theory will have much that Grice would recognize and appreciate. However, I offer several criticisms of the elements of Grice's work, and by combining his separate papers on ethics and psychology, I have created something new and unique.

The second major aim is to explore the relationship between this theory of moral psychology and Grice's theory of implicature. The picture that will emerge will simultaneously derive much of Grice's theory of implicature from Rational Virtue Theory while also modifying and correcting it in various ways. As we will see, Grice's Cooperative Principle (or an appropriately modified version) is a highly normative notion whose rationale stems from it being in one's self interest to be cooperative. This normative link will entail several changes and corrections to Grice's theory of implicature. The range and type of maxims that are at work in the communicating and understanding of implicature are much wider than Grice supposed.

The project consists of six chapters. Chapter 2 is a short review of Grice's theory of implicature, where I highlight several unexplicated notions – such as rationality and cooperation –

that warrant further discussion. Much of what follows will work to fill out these notions and consider what theoretical changes are needed as a result.

In chapter 3, I present a critical review of the literature left by Grice on philosophical psychology and happiness. Grice suggests that intelligent and rational creatures who can set their own ends are capable of producing a manual, which contains rules for behavior. If followed, these rules should maximize these creatures' likelihood of realizing their ends. Grice turns to happiness to provide guidance as to what ends are rational. Happiness, he says, is an inclusive end, i.e., a set with many ends as members. To aim at happiness is to select ends from that set (which becomes one's personal happiness set) and then to work towards realizing them. The coordination of ends in a personal happiness set is also critical.

In chapter 4, I sketch such a manual. According to the theory that emerges, it is rational to have certain ends and to possess certain virtues, for doing so is the optimal means for achieving one's goals and being happy. The manual contains *ceteris paribus* behavioral principles. It also contains rules for the adoption of final and instrumental ends, as well as rules for the balancing of those ends. The chapter concludes with brief responses to the challenges of egoism and situationism.

Chapter 5 focuses on cooperation, both in general and in language use, arguing that it is a *ceteris paribus* behavioral principle included in the manual. I begin by reviewing the work of Michael Tomasello in evolutionary anthropology and developmental psychology. He compares cooperative behavior demonstrated by human children and chimpanzees and reaches a conclusion about the origin of cooperation in human evolution. I then turn to Christina Bicchieri's work on

social norms and Brian Skyrms work in evolutionary game theory. They both present game-theoretic arguments to justify cooperation as rational behavior. Cooperation in language, I contend, is fairly distinct. As I will explain later, we will not be able to use the same means to justify cooperation in conversations as Skyrms and Bicchieri used for cooperation in general. Analysis of the problem prompts a substantial revision of Grice's Cooperative Principle. I argue that there are two distinct notions of cooperation in language use. The one Grice opted for is neither the source of his conversational maxims nor rational to conform to. However, my new version – Communicative Cooperation – is both rational and the actual source of the Gricean maxims.

In chapter 6, I apply this connection between moral psychology and language to Grice's theories of meaning and implicature. I argue that moral maxims can play the same role in generating non-conventional implicatures as Grice's conversational maxims do. This insight spawns a new taxonomy of implicatures and a deeper understanding of their generation. Additionally, I argue that the Cooperative Principle is just one of several Conversational Principles at work in guiding our utterances and interpreting the utterances of others. Finally, I demonstrate how some of Grice's critics have incorrectly attributed various errors to his theory of implicature, instead of to speakers who actually make the mistakes. The existence of an implicature and the speaker's success in communicating the intended meaning are two separate issues.

In chapter 7 I attempt to dissolve the is-ought problem by making explicit a previously unrecognized implicature in the debate. It has been argued (perhaps first by Hume) that one can't derive an 'ought' statement from an 'is' statement. I argue that hidden in this argument is a normative implicature that one ought not try such a derivation, that is, that one should not

commit the is-ought fallacy. This 'ought' implicature, however, is built upon an 'is' statement itself.

In chapter 8, I review what has been accomplished, examine how this project sets the stage for a future virtue-theoretic semantics. These cases are very similar in terms of how the deception is achieved, but would seem to be of distinct moral character.

Chapter 2: Grice's Theory of Implicature

I will begin with an overview of Grice's theory of implicature, primarily as articulated in his (1975a), highlighting several concepts Grice introduces without much explication.

For Grice, the most basic notion of meaning is that of what an utterer *S* means by uttering a sentence *x* on a given occasion. What the utterer means is a function of certain (rather complex) audience-directed intentions he or she had in making that utterance. (Grice 1957, 1969, and Schiffer 1972). The precise details need not concern us here. What is important is that the notion of intention is central to the other semantic notions Grice defines in terms of this basic notion of meaning.

Not everything an utterer means is conveyed directly by the words he uses. And this fact motivates Grice's first major distinction between meaning and saying. What someone has *said* (in Grice's sense of the word) is "closely related to the conventional meaning of the words (the

sentence) he has uttered" (1975, pg. 44). But what someone has said, on a given occasion, by uttering some sentence *x*, need not be identical to what he *meant* by uttering *x*. The principle reason for this is that speakers can *implicate* things they do not literally say, e.g. they may, *imply, suggest, insinuate* or *hint at* things they do not actually say. speaker's meaning. In this gap between what is said and what is meant, Grice interjects implicature, which is what the audience can take the speaker to have communicated beyond (or instead of) what was said, on the assumption that the speaker is rational. This rationality requirement is a key component that most commentators on Grice pay lip service to and quickly move on. By explicating the nature and justification for rationality underlying Grice's theory of implicature, I will expand what fills this gap between saying and meaning to include additional types of implicature beyond what Grice discussed.

For Grice, the main type of implicature is conversational implicature. How he lays the groundwork for this notion is significant. He states,

> Our talk exchanges do not normally consist of a succession of disconnected remarks, and would not he *rational* if they did. They are characteristically, to some degree at least, *cooperative* efforts; and each participant recognizes in them, to some extent, a common *purpose* or set of purposes, or at least a mutually accepted direction. (1975a, pg. 45; emphasis mine)

In order to be rational conversational participants, Grice expects them to observe his Cooperative Principle: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (1975a, pg. 45). Grice then presents four conversational maxims, the following of which will

generally result in adherence to the Cooperative Principle: Quantity, Quality, Relation, and Manner. Grice then characterizes conversational implicature as follows:

> A man who, by (in, when) saying (or making as if to say) that *p* has implicated that *q*, may be said to have conversationally implicated that *q*, provided that (1) he is presumed to be observing the conversational maxims, or at least the Cooperative Principle; (2) the supposition that he is aware that, or think that, *q* is required in order to make his saying or making as if to say *p* (or doing so in THOSE terms) consistent with this presumption; and (3) the speaker thinks (and would expect the hearer to think that the speaker thinks) that it is within the competence of the hearer to work out, or grasp intuitively, that the supposition mentioned in (2) is required (1975a, pg. 49-50).

What a speaker has conversationally implicated then must be able to be worked out by the audience based on: (1) "the conventional meaning of the words used, together with the identity of any references that may be involved;" (2) the conversational maxims and the Cooperative Principle; (3) the context of the utterance; (4) other background information; (5) the supposition that all facts from (1) – (4) are available to all participants of the conversation (1975a, pg. 50). There is a problem here in that in working out what is implicated, the hearer begins with what the speaker said, *p*, and assumptions (1) – (5) listed above as premises. What is conversationally implicated, *q*, is not in any premise and cannot be logically derived or calculated from the premises alone.

It is worth noting a few other comments by Grice on rationality. "As one of my avowed aims is to see *talking as a special case or variety of purposive*, indeed *rational*, behavior, it may be worth

noting that the specific expectations or presumptions connected with at least some of the foregoing maxims have their analogues in the sphere of transactions that are not talk exchanges" (1975a, pg. 47; emphasis mine). Later he states,

> So I would like to be able to show that observance of the Cooperative Principle and the maxims is *reasonable* (*rational*) along the following lines: that anyone who *cares* about the *goals* that central to conversational/communication (such as giving and receiving information, influencing and being influenced by others) must be *expected* to have an interest, given suitable circumstances, in participations in talk exchanges that will be *profitable* only on the assumption that they are conducted in general accordance with the Cooperative Principle and the maxims" (1975a, pg. 49; emphasis mine).

Grice's theory assumes an unexplicated notion of rationality. He also incorporates notions of cooperation, purposes or goals of conversations, profitability, and calculability. None of these are explained in great detail. I am not objecting to these assumptions, only noting their presence and the need for further exposition. When we expand our attention to Grice's other works, we discover that he held views on many of these points, though not always fully worked out or connected together. In what follows, I will develop a theory of rationality and moral psychology along Grice's suggestions that will ground the conceptions of cooperation and goals in conversation. However, in so doing, I will argue that various elements of Grice's theory of implicature must be re-examined and modified.

Chapter 3: Gricean Foundations in Ethics and Psychology

To date there is no published work in which Grice presented his entire moral psychology. Instead we have separate writings on the related subjects of philosophical psychology, happiness, and ends. This chapter will serve as an exegetical review of Grice's relevant writings, from which Grice's moral psychology can be extrapolated, which will lay the groundwork for me to construct my own theory of moral psychology along Gricean lines, which I will take up in the next chapter.

The two papers of Grice's I will focus on are his 1975 APA Presidential Address "Method in Philosophical Psychology (From the Banal to the Bizarre)" (hereafter "Method") and his "Some Reflections about Ends and Happiness" published in 2010 in *Aspects of Reason* (hereafter "Reflections"). I shall also draw on concomitant remarks by Grice scattered throughout his other writings and the commentary and notes by his students Richard Grady and Richard Warner.

It will be helpful to begin by getting a sense of Grice's general approach to philosophical problems and methodology he employs in these articles. First, Grice does not claim to base any of his work on empirical findings. He claims, rather, to be developing ideas presented (sometimes obscurely) by Aristotle, Kant, and Hume by working out conceptually clear and consistent definitions of psychological notions that could form the basis of an explanatory psychology.

Second, for Grice, the concept of *value* is an integral part of rationality (1989, pg. 298). To be rational, for Grice, is to *evaluate*. He also speculates that naturalistic attempts to characterize rationality will inevitably fail. For Grice, value is a fundamental and non-natural component of reality, and specifically of rationality.[1] Any adequate theory of psychology must account for the fact that people have values and are evaluators.

Third, in his discussions of psychology and ethics, Grice employs the same methodology he used in addressing a host of philosophical questions, a methodology best known from his discussion of conversational implicature. In undertaking to explain someone's behavior or psychology, the strategy is to provide an explicit reasoning to that behavior, belief, etc. If such an account can be provided, then the *explanandum* is rationally justified. Grice did not claim (and was not worried) that people do not typically reason in the way he described. It would be preposterous to propose that people typically work out conversational implicatures by reference to Grice's

---

[1] Grice does not offer arguments for this non-naturalistic position. For my part, I do not see the position as essential to most of his views on value (or rationality). It is not a position with which I agree and the Gricean theory I produce in subsequent sections will not be committed to the view. It would seem that if mental states like believing, valuing, etc. are supervenient upon and functionally reducible to physical body states, then this non-naturalistic picture can be avoided. Grice, however, appears highly skeptical of physicalism in general and the possibility of such functional reduction of the mental to the physical.

conversation maxims and engaging in an instance of his "working out schema." The point is that one could *in principle* so reason.

1 Philosophical Psychology

In "Method," Grice sets out his methodology for philosophical analysis of a "battery" of concepts from which a psychological theory can be built. The purpose of a psychological theory is to explain behavior systematically and in a way differing from any presupposed theory aiming at the same goal, such as a physiological theory. The psychological theory is to be understood through an understanding of the theory's concepts. Psychological "laws or quasi-laws" make use of these concepts, and we can explain the concepts through their roles in those laws, while the laws explain behavior (Grice, 1975b, pg. 26). Psychological laws are not, however, strict causal laws but *ceteris paribus* laws, similar to probabilistic generalizations without specific weights assigned to them. Grice argues for the legitimacy of any posited psychological laws and concepts by means of the liberal and paradigmatically pragmatic criterion he dubbed "ontological Marxism" – "*they work therefore they exist*" (Grice, 1975b, pg. 31). We are justified in accepting into our psychological theory any law that sufficiently accomplishes the theoretical work required of it, namely to explain some class of behaviors.

1.1 The Methodology of Creature Construction

Grice is impressed by the fact that the same psychological concepts are attributable to different types of creatures of increasing psychological complexity, with human beings (currently) standing at the pinnacle.[2] The assumption of creatures related via a hierarchy of complexity serves as a simplifying mechanism for Grice: with each increase in complexity, work need not begin from scratch in developing psychological concepts for that type of creature. One needs only take the concepts required by a slightly less complex creature and supplement them with a few more laws or concepts.

Working from this point, Grice developed a methodology of creature-construction for positing and justifying psychological concepts. We can, he claims, place ourselves in the role of genitors, whose purpose is to design a type of creature (in our imagination), asking what psychological concepts and functions such creatures will need in order to survive.[3] Such creatures we can call 'pirots,' and they can increase in complexity as we desire.[4] The procedure of constructing pirots is primarily elucidating when the type of pirots we are designing matches (at least closely enough) actual creatures. For instance, our pirots can be "squarrels," which are

---

[2] Whether this is a scientific fact or a pre-theoretic intuition on Grice's part is perhaps debatable. For my own part, I think that he is likely correct, but it is not immediately relevant to work on human moral psychology. Because *ex hypothesi* humans are the only type of creature (currently) possessing a moral psychology, we need not refer back to any less complex species to develop a philosophical account of this aspect of our psychology.

[3] That this is type, not token creature construction will be relevant in Section 2 when developing a moral psychology.

[4] Grice borrows the term 'pirot' from Carnap (1937) and Russell (source unknown). Carnap appears to have made up the term so that he could discuss syntax and deduction. He tells us that "Pirots karulize elatically." If we also know "A is a Pirot," then we can deduce "A karulizes elatically" without knowing what a Pirot is and what it means for something to karulize elatically.

sufficiently similar to squirrels. So, we learn about the psychology of squirrels by constructing squarrels.[5]

In placing ourselves in the genitorial position for a given type of pirot, we assume certain operational facts, for example, operational energy requirements and the required frequency of certain operations, such as energy consumption. Grice does not discuss a host of other biological facts (or operational facts akin to biological ones) that must be assumed in order to ensure a correspondence with real-world species. For instance, we must also assume the presence and function of various biological structures, such as organs, and what sort of activities or conditions would be fatal to the pirot. We can assume all these and other such relevant facts are given to the genitors.[6]

With these operational characteristics provided, Grice establishes three constraints upon the genitors.

(1) First and foremost, they are to design the pirots with an eye toward their survival, i.e., the survival of each individual pirot.[7] There are two reasons for this genitorial goal. There seems

---

[5] To be clear, a type of pirot corresponds with a species.

[6] A potential criticism of the methodology of pirot construction is that it does not differentiate between sub-categories within a given type of pirot, such as sexual differences for instance. Grice's aim is to develop psychological explanations for all pirots of a given type, based on the assumption that for any behavior that all such pirots perform, the psychological explanation will be more or less the same, typically regardless of other differences between the pirots (unless there are compelling reasons not to). We can then apply this methodology to any behavioral difference among a type of pirot that tracks with sub-categories.

[7] The survival of the species is another matter, which though important, is not immediately relevant. We will come to that in due course. As a preview, the survival of the individual and the species are not in complete opposition. The species survives because enough of the individuals survive (and procreate).

little point for the genitors to create them otherwise. And more importantly, there is no reason to believe that there would be any actual species to match these pirots otherwise, for such non-survival-oriented creatures will not actually exist (for long, anyway). This is an important point that all pirots, regardless of type, have survival fundamentally built in as a goal. The genitor's aim then is to provide the pirot with a psychology that optimizes its chances of survival.

(2) The second genitorial constraint requires reference to the pirots' living conditions must be very general, so that they might survive in wide-ranging environments. While in general this is a good point, it is open to several points of criticism. It is likely too broad at least for most species. There seems little reason to think that a squirrel's psychology is general enough (i.e., in terms of reference to its living conditions) for it to survive in frozen Antarctica. We cannot let our squarrels get too far beyond squirrels in terms of survivability in different environments. It seems probable that more psychologically complex and intelligent species will be able to survive in a wider range of environments. Thus, how general the genitor's reference to living conditions is may well depend upon the complexity of the pirots in question, which means that the degree to which a pirot is capable of psychological complexity must be another assumed biological fact from the outset.[8] The second constraint may be too broad in another way. Because the genitor wants squarrels to survive in a wide range of environments, he would talk of squarrels' desire for food to explain the eating of nuts. This would then require a psychological mechanism for the squarrel to

[8] It is a legitimate question to ask why such simple creatures have any psychology whatsoever. Grice simply assumed they do. There is then also the question of psychological threshold. Do all organisms have at least some psychology, and if not, then at what point (biologically, we assume) does psychology present itself? Grice stated that plants will not require any psychological concepts to explain their "behavior," but offered no explanation of why not or what the threshold is.

recognize nuts as food. For actual squirrels, it seems more likely that they have the much more narrow concept of nuts, instead of food in general.[9]

(3) The third genitorial constraint is parsimony: genitors should not ascribe any more psychological concepts to pirots than is explanatorily necessary. While *prima facie*, this would seem a good restriction (and probably right most of the time), there are reasons that suggest parsimony of psychological concepts and laws is not typically observed by nature. Evolution is not elegant and does not observe parsimony. What we could conceptually explain with one psychological law may have developed in pieces at different points of a species evolutionary history, so that in fact several different psychological laws are at work. Additionally, it seems intuitively plausible that one species is slightly more psychologically complex and intelligent than another precisely because the more intelligent species has developed a more efficient psychology with fewer concepts or laws that do just as much work (or even more with less). So, not every step up in complexity requires a psychological addition. That said, however, parsimony of psychological concepts and laws is *ceteris paribus* methodologically sound. Evolution may have provided a species more psychological mechanisms than are strictly required to survive. But any extra psychological machinery should be empirically observable (either by behavioral differences or in brain activity, such as with an fMRI), and thus fall under what must be parsimoniously explained.

Following from parsimony, Grice states, "no psychological concept can be instantiated by a pirot without the supposition of behavior which manifests it" (1975b, pg. 39). The only point of

[9] The combination of the second, third, and fourth constraints (see below) might suggest that the genitor should construct squarrels in a similar way (that is, to desire nuts instead of food in general). However, it is not immediately clear that this is required or that Grice considered such factors.

criticism I have here is that Grice does not clarify whether a behavior must have previously been manifested or merely be manifestable by a pirot. For example, perhaps there is some behavior that squirrels are capable of manifesting – karulizing for example – but it just so happens that the conditions have never actually arisen in which any squirrel would karulize. Must a psychological theory for them account for this potential behavior, or may we simply limit ourselves to all the behaviors squirrels actually have demonstrated? Grice does not give a clear answer to this question. I believe the right answer to be manifested behavior; otherwise anything could be possible.

1.2 Self-Genitorial Position and the Manual

Despite the points of criticism raised above, Grice's pirot construction procedure is not fundamentally flawed, but rather under-developed. I take the method of pirot construction to be a highly useful tool for philosophical psychology, particularly when employed in conjunction with other methodologies. At present, however, the real value of pirot construction comes when we design especially complex pirots similar to human beings.

1.2.1 Grice's Remarks

The genitorial position is, of course, a fiction (though one that Grice thinks can be recast in terms of final causes). What is interesting for our purposes is that actual creatures of sufficient complexity that are very intelligent and rational can place themselves in the genitorial position, i.e. they can decide how they would design themselves with a view to their own survival.[10] Describing

[10] Again, we note that self-genitorial pirots, in entering the genitorial position, engage in creature-*type* construction, not creature-*token* construction. A pirot of this type is not here/yet asking how to maximize the chances for his or her *own* survival; that will come later.

creatures of this type, Grice calls them "very intelligent rational pirots" (1975b, pg. 40). By describing pirots of sufficient complexity, intelligence, and rationality, we make them such that they can redesign themselves; they can set their own ends. We may call them 'Gricean pirots.'

When building simpler pirots, the genitors provided them with certain ends as built-in for the sake of survival that can't be modified by the pirots. These permanent and built-in ends make them more likely to survive. A pirot, by working to satisfy these ends was thereby maximizing its chances of survival. But Gricean pirots are special. "These pirots will be capable of putting themselves in the genitorial position, of asking how, if they were constructing themselves with a view to their own survival, they would execute this task" (Grice, 1975b, pg. 40). A Gricean pirot who places itself in the genitorial position will first be able to understand the reason its genitors designed it the way they did. Being able (as Gricean pirots) to redesign themselves is a highly efficient survival mechanism, which a Gricean pirot will comprehend in the genitorial position. [11]

The second thing a Gricean pirot can do upon entering the genitorial position is author "a highly general practical manual," which helps all pirots of that type in their bid for survival (Grice, 1975b, pg. 40). Grice's conjecture is that the contents of this manual, if heeded, will maximize the chances of survival for these pirots. But far from being a code of conduct handed down from on

---

[11] Grice asserts that a Gricean pirot will be able to justify its own continued existence, but does not elaborate much on this claim, which deserves further examination. Some of the views espoused in *The Conception of Value* (where "Method" was also reprinted as a chapter) suggest that Grice thought that the reason one is able to justify one's continued existence relies upon a potentially controversial view he developed called "Metaphysical Transubstantiation," whereby human beings, who have rationality as an accidental property, recast themselves as persons, for whom rationality is an essential property. At present, I have no desire or intention to develop or make use of this metaphysical view, but the question of how Gricean pirots justify their continued survival is one that deserves further consideration. My extension upon Grice presented in Section 2 should go some way to that end.

high (either divine or metaphysical), Gricean pirots compose this manual for themselves precisely because it is in their best interests. Grice insists that the contents of the manual must be *universalizable*. The genitorial position, as we noted previously, is concerned with pirot-type construction and does not focus upon designing individual pirots. Each pirot of that type operates according to the psychological laws and by means of the psychological concepts established for the entire type. So, in a way similar to Rawls, by entering the genitorial position, a Gricean pirot cannot by definition write a manual solely for itself.[12] A significant problem is that Grice does not argue why following this type manual increases the chances of survival for an individual Gricean pirot, instead of (at least occasionally) going against it. This hole will become the challenge of the egoist to Grice's view, and I will address the question in the next chapter.

Besides universalizability of content, the manual must be general in nature; Grice identified three necessary types of generality, though he offered little detail on each. The manual must have *conceptual generality*, since it can only be composed of the concepts that have already been genitorially justified. The manual cannot, he argues, refer to any sub-class of Gricean pirots. Hence, the manual must have a *generality of form*. Finally, Grice states that a Gricean pirot in the genitorial position would not include content that is only applicable in circumstances to which that pirot cannot ever be subject. It must be the case that any pirot could in principle from time to time find itself in the relevant circumstances to heed any injunction in the manual. Thus, the manual must have *generality of application*.

---

[12] In *A Theory of Justice*, pg. 429ff. Rawls addressed moral psychology in relation to a sense of justice. His brief discussion psychological laws, ends, and intentions on pg. 433 bears some resemblance to Grice's work and the picture of moral psychology that will emerge in Section 2.

Finally, though the manual is both general and universal, according to Grice it is permissible to use the manual as a basis for "more specialized manuals to be composed when the pirots have been diversified by the harsh realities of actual existence" (1975b, pg. 40). I take it that Grice thought of these specialized manuals to be composed only by additions to the general manual, so that no general injunction is removed for a sub-class of pirots. The generality of these specialized manuals will vary inversely to how small each sub-class is: the smaller a sub-class of pirots, the more specific the specialized manual is for them. At this point, the Kantian overtones should be abundantly clear, so that it should be of no surprise that Grice titled the manual 'IMMANUEL.' Human beings, he thought, were the sort of "very intelligent rational pirots" he had been describing, who compose IMMANUEL and sometimes heed it "(in our better moments, of course)" (1975b, pg. 41).

1.2.2 Comments from Richards

Since this is the extent of Grice's comments both in "Method" and elsewhere in his published works on the manual, it is worthwhile to turn to discussion by his students and commentators, Richard Grandy and Richard Warner (collectively referred to as 'Richards' by Grice and in what follows) to fill in more of the picture. Though we need not assume their commentary is a completely accurate presentation of Grice's views, it is useful to consider in developing a Gricean moral psychology. They first make a helpful clarification. For any pirot, regardless of complexity, the genitors build in certain ends, which function as a constant end that the pirot endeavors to satisfy in all suitable circumstances (Grandy & Warner, 1986, pg. 31). By providing pirots with a multiplicity of ends, genitors design them to schedule their endeavors so that they are not trying to satisfy all of them all the time, but so that at any time they do have some

built-in end to work towards. Additionally, how they schedule their efforts should be in whatever way maximizes their chances of survival.

Life for Gricean pirots, however, is not just a matter of survival. Richards suggest that Gricean pirots are given the built-in end of being end-setters (Richards' term), that is, of being pirots that can modify their ends. Life for them then becomes a matter of maximizing their chances of satisfying their built-in ends. Thriving, not just surviving, is the aim of Gricean pirots.[13] They, as end-setters, do not have the capacity to change any and every end; some of the ends built-in by their genitors cannot be changed, namely the end of being an end-setter. Yet, they can change at will most of their ends, and they do so for the sake of maximizing their chances of satisfying their built-in ends. (Presumably, this means satisfying those ends that cannot be changed or eliminated, though Richards are not clear on this point.) Applying this notion of built-in ends back to Grice's comments, it would seem that Gricean pirots are at least able to understand the reason for and benefit of the built-in ends. But for pirots to know what other ends to add, change, or eliminate (and under what conditions) requires that the genitors provide the pirots with evaluative principles. These principles represent the wisdom of the genitors, such that the genitors need not continually redesign Gricean pirots; they are able to do so themselves to the same effect. The term 'evaluative principles' is an addition of Richards to the Gricean picture, though they do not clarify whether they are part of the manual or the wisdom Gricean pirots use when either

[13] Grice and Richards continue to talk about the survival of Gricean pirots, though in a way distinct from the survival of lower pirots, in an apparent redefinition of the term. I will avoid this terminological confusion by shifting to thriving.

composing or referring back to the manual.[14] Instead of undertaking here the lengthy task of criticizing and responding to Richards in sufficient detail, I will simply recast Gricean pirots in the next chapter after reviewing Grice's work on happiness, which will elucidate more of the contents of IMMANUEL, what ends can be modified, and the place for the evaluative principles.

The method of pirot construction represents a new approach in philosophical psychology. It does not take as a starting point our linguistic intuitions about psychological states and then seek to systemize them. This is a common methodology currently employed by many experimental philosophers. Grice is not adverse to this approach, but appears to think the methodology of creature construction will "tell me sooner and louder if I am on the right track" (Grice, 1975b, pg. 36). For my part, I prefer a marriage of those two methodologies, thereby allowing for our attempts at creature construction to be better empirically informed.

Finally, Warner notes that his methodology was regularly discussed in Berkley in the 1970s and related the following story. Once, Grice was directly challenged on the legitimacy of this method. With exasperation, he responded, "But there must be a rational explanation" (Warner, 2001, pg. *x*). The explanation establishes its rationality. The claim will not be that people actually do reason to moral principles in the way discussed. But because a rational explanation is possible, conforming to the proposed moral principles will be rational and normatively justified.

2 Ends and Happiness

[14] Grice spoke of "a set of criteria for evaluating and ordering ends" (1975b, pg. 40). This may have been the same idea Richards referred to with 'evaluative principles.' While they have done a helpful service to Grice on this point, much of their explication further muddled the picture of the manual.

The second paper to which we now turn is Grice's "Some Reflections about Ends and Happiness" from his book *Aspects of Reason* (2001). In "Reflections" Grice is working towards an analysis of desire and its psychological roles in rational creatures. To limit the project, he relates this discussion to various questions arising from *Nicomachean Ethics*. Given the focus on rationality and psychology above, "Method" and "Reflection" have a clear connection, though not one publicly presented by Grice in his lifetime or in publications to date.

In "Reflections" Grice's main focus is to advance the following four theses, each of which I will explain in turn. The first is:

1) Happiness is desirable for its own sake, and other ends (desirable for their own sakes) are also desirable for the sake of happiness, which is an inclusive rather than dominant end.

Before specifying the dominant/inclusive distinction, we first must begin with the concepts of maximal finality and self-sufficiency. Grice notes that for Aristotle, whatever it is that is good for a person must meet two conditions, maximal finality and self-sufficiency. 'Maximal finality' is Grice's own term for the following Aristotlean notion:

> Now we call that which is in itself worthy of pursuit more final than that which is worthy of pursuit for the sake of something else, and that which is never desirable for the sake of something else more final than the things that are desirable both in themselves and for the sake of that other thing, and therefore we call final without qualification that which is always desirable in itself and never for the sake of something else. (*Nicomachean Ethics*, 1.7)[15]

[15] Grice used Ross's translation (1925), which is given here.

Grice followed Aristotle on self-sufficiency, defined thus: "The self-sufficient we now define as that which when isolated makes life desirable and lacking in nothing" (*Nicomachean Ethics*, 1.7). *Eudaimonia* is thought to fulfill both requirements. [16]

Grice objects to the view of many of Aristotle's commentators, who interpret Aristotle as claiming that happiness is the one and only end that is desirable for its own sake. Rather, Grice contends that Aristotle held (and was right to do so) that there are a number of ends (or desires),[17] each of which is desirable for its own sake. As an example, Grice suggests that lying about in the sun is desirable for its own sake, and so has maximal finality.[18] Consequently, *eudaimonia* is merely one such end.[19] To clarify his point, Grice distinguishes between ends that are desirable for their own sake (which he called 'I-desirables') and ends desirable for the sake of *eudaimonia* (which he

---

[16] Grice initially is non-committal on whether the English word 'happiness' sufficiently captures the notion of *eudaimonia*. However, he spent most of "Reflections" referring to happiness and his student and commentator Richard Warner does the same. So even if there is some difference between the two concepts, Grice's interest is happiness. As Warner (1986) noted, Richard Kraut provides an argument ("Two Conceptions of Happiness," *The Philosophical Review*, vol. 88, no. 3, April 1979) that this is a good translation. I am *somewhat* skeptical on this point, but for the sake of this project will assume 'happiness' is good enough and will refer to it.

[17] In commenting on "Reflections," Richard Warner noted that there are some (un-named) philosophers who believe that we can only talk appropriately here in terms of ends, not desires, for one may have an end E without desiring E. Warner preferred to avoid this debate, apparently thinking little rides on the distinction. It is not clear from "Reflections" that Grice considered the matter. The objection is correct: ends aren't the same as desires and we should focus on ends. In the next chapter, I will offer a simple solution to the problem, which will legitimize all talk of desires (instead of ends) in this chapter.

[18] Aristotle stated that every virtue was desirable for its own sake, such that even if nothing resulted from having the virtues, we should still desire them; see 1.7, 1097b2-4.

[19] Obviously, this is a highly debatable point on Aristotlean exegesis. However, regardless of Aristotle's view, for this sake of this dissertation we can attribute this view to Grice's theory of happiness and move on. Though initially very skeptical, after a close re-reading of *Nicomachean Ethics*, I think there may be something to Grice's reading of Aristotle, but save such considerations for a future work.

termed 'H-desirables'). The notion of I-desirability is insightful, but not sufficiently developed by Grice. He claims that all H-desirables are I-desirables, but not all I-desirables are H-desirables. However, he offers no examples of I-desirable ends that are not H-desirable, so that it is not immediately clear what such ends might be.[20] His reason for allowing I-but-not-H-desirables are less than clear. Nevertheless, sense can be made of it, and the distinction between H-desirables and I-desirables will be important below in developing a moral psychology. As an additional point of criticism, there appears to be a tension in the text regarding whether an I-desirable end is desirable for its own sake for everyone or different people might find different ends I-desirable. The latter obviously has the intuitive appeal of common sense. Some find playing golf to be desirable for its own sake, though I do not. Grice sometimes speaks of the desirability of a life as solely determined by the one whose life it is. However, there are also passages with a decidedly more universal tone. In the next section I will work to resolve this tension.

The second thesis is about the nature of this set of ends that Grice regarded as *eudaimonia*. It states:

2) Happiness is a rational inclusive end, and as such its component ends must exemplify some open feature explaining their inclusion.

According to Grice, happiness is not a single, monolithic end, which he took to be the standard and incorrect interpretation of Aristotle.[21] Rather than construing *eudaimonia* as a "dominant"

[20] It is possible that the set of I-not-H-desirables may be an empty set because everything desirable for its own sake is desirable for the sake of happiness. This remains an open point for future debate, though nothing rides on it for the sake of this project.

[21] As a gloss for his concept of *eudaimonia*, Grice suggests that it is that state that one's good *daimon* or fairy godmother, concerned only for one's well being, would ensure for one.

end "consisting of just one valued activity or good" (Ackrill 1974, pg. 5), it is an "inclusive" end, i.e., a set of H-desirable ends.[22] Achieving *eudaimonia* depends on realizing the ends in this set.

But must one realize all H-desirable ends to be happy? That would be an impossible burden. Is one's life more *eudaimon* the more I-desirable (and H-desirable) ends one realizes? Grice thought not. There is some point after which there is a "vanishing marginal utility" for realizing any further H-desirable ends: it will not make one any happier. Sadly, Grice did not specify what the point is or describe it any way. Warner suggests that one is *eudaimon* realizing "enough" H-desirables (Warner, 1986, pg. 478-9). However, without establishing what constitutes enough, this is not as helpful a suggestion. I will offer some thoughts on this problem in the next section.

Grice stipulates that this happiness set **H** is not a mere aggregate of I-desirable ends, but rather has some organizing plan. To make his point he imagined a world W in which there are only three H-desirable ends (that are also I-desirable): A, B, and C. One might think then that for *x* to desire happiness, it is sufficient for x to desire the realization of A, B, and C. Not so, according to Grice. The set is defined intensionally rather than extensionally. Just because *x* desires A, B, and C, it does not follow that *x* wants them *for the sake of happiness*. They are desirable in themselves, and *x* may desire them only as I-desirables. Rather, Grice argues that *x* must want A, B, and C for the sake of happiness, which requires that "there is an 'open' feature F which is believed by *x* to be exemplified by the set" {A,B,C} (1975b, pg. 122). This open feature is supposed to explain membership in the happiness set {A,B,C} without enumerating its members. He explains this feature by way of example. There was at some time an Oxford don who desired to

[22] I follow Grice in quoting Ackrill here, who defined the terms 'dominant' and 'inclusive.'

teach at the colleges of Somerville, St. Hugo's, St. Hilda's, Lady Margaret Hall, and St. Anne's. Teaching at any of those colleges might well be desirable for its own sake, but his reason for wanting to teach at all five was that they were all the women's colleges at the university. To explain the nature of the don's desire, we need not list each college he wishes to teach at and why, but rather we need only refer to this feature of his compound desire. But not any F will do. For any non-empty set, there will always be some F, such that F is true of all members of that set and none other; if nothing else, that they are members of the set. I take it then that the open feature of the happiness set is supposed to capture the nature of the set as a universal. The crucial point here is that all and only those compound desires that have such an open feature are, according to Grice, *ipso facto* rational desires. He then assumes that happiness, as a compound desire, has some open feature.

There are a few points of criticism to offer on this second thesis. The first is that the open feature for happiness remains open; Grice says little as to what it might be. Whatever it is will determine which I-desirables are in fact H-desirables and which are not. It would seem crucial for a theory of happiness that claims happiness is an inclusive end to provide insight into what sort of ends it might include. The lack of explanation about the nature of this open feature of the happiness set is perhaps the biggest missing component of Grice's theory and point of greatest animadversion. Without it, Grice's moral psychology is incomplete. However, in the next chapter I will make some intuitively and experimentally well-informed assumptions about whether certain ends are included in happiness or not. Additionally, absent a full account of this open feature, we can still develop a moral psychology in outline, even if its full contents cannot yet be supplied. Yet Grice offers a hint about the open feature of happiness at the very close of "Reflections." He

thinks Aristotle was on the right track in appealing to human beings' essential character as rational animals. Hence, it might be possible to determine what sort of ends happiness includes by asking what ends a creature-constructor would endow in such rational animals if he were designing them with an eye towards maximum viability in human living conditions, but across a wide variety of environments. In the next chapter, I will explore some of these methods for partially filling in an account of happiness, which will be enough to derive the Cooperative Principle.

The second criticism is that while discussing the example of A, B, and C, Grice suggests without real argument that *x* could want A, want B, and want C without it being the case that *x* wants A, B, and C. This is a questionable position, and one to which I have objected elsewhere (Alfano & Robinson, unpublished manuscript). Desire is distributive. The third criticism is that Grice attempts to recast his explanation of I-desirables and of happiness in terms of universals. The end result is both unhelpful and unclear. Treating happiness as a set of ends and use of recent formal developments in a logic of desire should be clearer and will be the manner in which I proceed.

Grice's third thesis states:

3) The desire for happiness is a higher-order desire for a set of lower-order desires, each of which epitomizes this open feature.

For the third thesis, Grice examines the following story. Suppose that a tyrant punishes a court minister by assigning him the duties of attending to the palace's garbage, meaning to humiliate him, and threatens the minister with execution if the task is not performed well. Initially, the minister works diligently only to keep his head. But later, he realizes that by taking pride in his work, he can frustrate the tyrant's desire to humiliate him. So he then performs his duties for

their own sake in order to foil the tyrant. After many years of the minister taking pride in doing the menial task well, the tyrant is overthrown and the minister chooses to leave the job. One might say that the minister did his job efficiently both for its own sake *and* for the sake of spoiling the tyrant's plans, an apparently contradictory state. Grice finds this reading inadequate and explained it using higher-order desire. He points out that the tyrant's will would not have been frustrated merely by the minister doing a good job, but by the minister wanting to do a good job for its own sake. Only then will the minister take pride in his work and so not be humiliated. So the minister has the compound desire: he desires to do his job well (for its own sake) and he desires to so-desire for the sake of frustrating the tyrant. Grice thinks the same considerations apply to the desire for happiness. It is not enough to desire each I-desirable end in the happiness set. One must also desire that one desire them for the sake of happiness.

The notion of higher-order desires can be clarified by expressing the matter symbolically. The expression $D_x(\varphi|\psi)$ represents that an agent x desires the end φ given some condition ψ. For simplicity, we can assume that any I-desirable, such as A, is desirable in all circumstances.[23] So for some agent x, $D_x(A|T)$, where the truth-conditional T means x desires A always, regardless of circumstance. In world W, there are three I-desirables in the happiness set – $\mathbf{H} = \{A,B,C\}$. Agent x desires each I-desirable separately: $D_x(A|T)$, $D_x(B|T)$, and $D_x(C|T)$. From this we can conclude $D_x(A \wedge B \wedge C|T)$. Grice requires that x also have a second order: $D_x(D_x(A \wedge B \wedge C|T)|T)$. This expression represents x's desire (given that A, B, and C are all the members of the happiness set) to

[23] This assumption is permissible for now and in such a world as simple as W, with A, B, and C as the only I-desirables. In reality the matter is more complex, and one does not desire an I-desirable in all cases.

desire the three I-desirables. The desire for happiness (*Eu*) then is the compound of these two desires:

$$D_x(Eu \mid T) = D_x(A \wedge B \wedge C \mid T) \wedge D_x(D_x(A \wedge B \wedge C \mid T) \mid T)$$

Happiness is desirable for its own sake, and hence desired by *x* in all circumstances. To desire happiness means to desire all the I-desirables in the happiness set (or enough of them) and to desire to desire them because they are in that set.

The fourth thesis transcends the simplicity of world W and moves to the many complexities of the real world. It states:

4) One can construct for oneself a system of ends for the direction of life; some systems of ends are better than others by virtue of the fact that the adoption of that system of ends will *ceteris paribus* lead to realizing one's happiness; and there are criteria for evaluating systems of ends.

Happiness, for Grice, should serve to give direction to one's life. An understanding of happiness as a concept and as the end for human life should enable one to construct a system of ends for oneself that will serve as a guide to life. Grice's final project in "Reflections" is to offer criteria for the construction and evaluation of a person's system of ends.

He begins by distinguishing between happiness-in-general and happiness-for-a-person. We must extrapolate his view from his limited comments, but it appears to be that happiness-in-general is something akin to what we have called the happiness set. It is a set of I-desirable ends, which we can suppose must have some open feature to explain set membership. It contains all H-desirable, but not necessarily all I-desirables. Let's call the set of all I-desirables **I** and the set of all H-desirables (i.e., happiness-in-general) **H**. It follows that **H** is a proper subset of **I** ($\mathbf{H} \subset \mathbf{I}$). Grice's

conjecture is that an individual *x* can build his or her own system of ends exclusively out of **H**. There are too many H-desirable ends for any one person to satisfy all or even enough of them. So *x* chooses from the H-desirable ends in **H** based on *x*'s "character, abilities, and situation in the world" (1975b, pg. 131). The ends *x* selects are to function as a guide to life for *x*, and constitute happiness-for-*x*. We can call this personal happiness set for *x* $\mathbf{P_x}$, and it is a proper subset of **H** ($\mathbf{P_x} \subset \mathbf{H}$). Considerations from the previous theses still hold for $\mathbf{P_x}$. For $\mathbf{P_x}$ to be a rational compound end, there must be some open feature F_x that explains why *x* include some H-desirable ends, but excluded others. Also, one must realize the ends within it (or enough of them) and do so for the sake or realizing $\mathbf{P_x}$ in order to realize the end of individual happiness. All this appears to follow from what Grice says. However, there is a problem in that it is not immediately apparent that everyone does (or should) have the same set of I-desirable ends, **I**. Assuming they do not, **I** will need to be recast as the set of all ends one *might* find desirable for their own sake. Then there is a proper subset of **I** for each person composed of the ends he or she *does* find desirable, which we can call $\mathbf{I_x}$. It then must be the case that $\forall y ((y \in P_x) \leftrightarrow ((y \in H) \wedge (y \in I_x) \wedge (y \in \{z \mid F_x(z)\})))$, where $\{z \mid F_x(z)\}$ denotes the set of all elements *z* for which the assertion $F_x(z)$ is true.

Grice argues that any good system of ends a person builds ($\mathbf{P_x}$) should have two essential features.[24] A system of ends should have flexibility and maximal stability. There is an obvious tension between these two properties, though for his part, Grice emphasizes stability and tacked on flexibility as an after-thought. A stable system of ends is necessary to allow one to continue to strive towards its realization despite changes in external circumstances without modification.

[24] The evaluative terminology is intentional. Grice thinks some systems of ends are better than others vis-à-vis how well they enable one to realize those ends.

Nevertheless, no one's system of ends is (or likely should be) completely stable; sometimes modifications will be necessary. In those instances, flexibility is a boon to any system of ends, for it accommodates any need for modification of one's ends, but with the fewest changes necessary. A sufficiently stable, yet flexible system of ends will minimize the likelihood of crises and necessity of a near total replacement of one system of ends with another.

Grice then discusses seven features of systems of ends that he found to be conducive to stability and characteristic of happiness.

(1) *Feasibility*: One's system of ends must be "workable." Any skills, knowledge, competencies, etc. needed need to be naturally supplied or realistically obtainable.

(2) *Autonomy*: Grice finds it preferable that the less dependent one is on outside aid, the better. The greater reliance on factors outside of one's control, the less likely one is to realize one's system of ends.

(3) *Compatibility of component ends*: Grice correctly realizes that it is not realistic to demand no competition between one's ends. There does, however, need to be some balance in competition; only so much will be allowed. We cannot countenance a system of ends, for example, that includes being a life-long high school teacher and becoming a billionaire. Grice does not discuss how this balance is determined or achieved. It would seem that it must be empirically informed. Additionally, it seems likely that some individuals will be capable of realizing a system of ends with a higher degree of competition than other people.

(4) *Comprehensiveness*: The more practical questions that one can answer by reference to one's system of ends, the better and more comprehensive that system is. A system of

ends that leaves a person in a quandary about what career path to choose is insufficient. That said, however, Grice states that it would typically be inappropriate to appeal to one's system of ends to answer practical questions like what one should have for dinner tonight. He is not clear on this point. I take it that he thought there to be a *ceteris paribus* threshold of simplicity, and questions below this threshold need not refer to one's system of ends to answer. For example, my system of ends does tell me that I should eat enough calories today. But there are simpler ways of answering whether to have Mexican or Thai for dinner than considering my life-scheme.

(5) *Supportiveness of component ends*: Pursuing some ends enhances our ability to pursue others. Not all ends need to be supportive in this manner, but the more that are, the more stable a system of ends becomes.

(6) *Simplicity*: The answers that a system provides to practical questions need to be simple and easy to understand.

(7) *Agreeableness*: Grice is very unclear on this point, but his idea appears to be the following. Realizing an end is *ceteris paribus* agreeable, and in different forms. Some forms of agreeableness (he does not specify which) regularly attach themselves to realizing ends. Such regular forms of agreeableness cannot be used to determine which system of ends is preferable over the other. However, there are other, less common forms of agreeableness (such as delight, Grice says) which due to their rarity can settle that question. If one system of ends is more liable to produce delight than another, then it is *ceteris paribus* the more preferable of the two.

I will close this section with some comments on this list. First, Grice's comments on agreeableness are cryptic at best. Our previous considerations about **I** and $\mathbf{I}_x$ achieve some similar result to what I take Grice to have been striving towards, and so I will avoid future reference to this feature. Further, it is not apparent why (3) and (5) – compatibility and supportiveness of component ends – need to be separated. I take it that much the same effect can be obtained by combining them into just *harmony of ends*: competition among ends should be minimized and they should support one another as much as possible.

There are also at least two necessary features absent from this list. First, a system of ends should be *evaluable*. A person needs to be able to determine (and the easier, the better) how well he or she is doing in working towards realizing their ends. Second, the ends should be *able to be ranked*. If we accept that one can be happy if one realizes enough of one's ends (whatever that might mean), then one ought to be able to arrange them by order of importance. Then one can know which ends are more important and which can be left unrealized.

3 Conclusion

The original version of "Reflections," when Grice delivers it as a lecture ended with the following. "But I have now almost exactly reached the beginning of the paper which, till recently, you thought I was going to read to you tonight, on the derivability of ethical principles. It is a pity that I have used up my time." Grice was working towards a more complete ethical theory, in which ethical principles can be derived from an account of happiness. Happily, we are not out of time and can continue where he left off.

Chapter 4: Rational Virtue Theory

So far, I have laid a foundation for a new theory of moral psychology by reviewing Grice's suggestive contributions. But they came in two separate pieces, not as a unified theory. There was also much that was right and intriguing by Grice's proposals. But, as we saw, his views were not without problems. Now it is time to move beyond Grice, building on what he left, but establishing something new and quite distinct from what Grice's moral psychology would have been, if he'd put his pieces together.

To continue where Grice left off, we enter the genitorial position for ourselves and compose IMMANUEL, the manual intended to maximize our chances of satisfying our self-determined ends. The primary project of this chapter will be to begin the composition of the manual and consider means for filling it in. That said, however, compiling this manual in its entirety will be a herculean task beyond this dissertation. I aim to provide some of its outlines,

some of its content, and some methods for composing it. For lack of a title from Grice, I shall dub the view that emerges 'Rational Virtue Theory.'

Like Gaul, the manual is divided into three parts: First, the manual needs rules for choosing ends, which includes rules both for the initial selection of one's ends and for any subsequent changes; Second, the manual must have some rules for organizing those diverse ends into some sort of guide for life; Third, the manual will contain behavioral principles for maximizing one's chances of satisfying those ends. Before proceeding though, the notion of rules warrants some discussion. I do not mean to suggest a deontic system. Rational virtue theory is far from an absolutist view of ethics. Typically, these rules are (or should be) *ceteris paribus* laws in our moral psychology. They also don't prescribe actions; that job is left to behavioral principles. My notion of rules will be refined as we progress.

Aside from developing these three components of the manual, this chapter will also respond to two objections to the theory of moral psychology that emerges. I do not endeavor to respond to all possible objections or to claim a fully worked out and defended view. The objections considered in the last section of this chapter are chosen either to further explain the emerging view or respond briefly to some major criticisms against virtue theories of ethics.

1 Selection of Ends

The permissibility of all ends is derived from their pragmatic value. That is, what sorts of goals are rational and permissible will entirely depend upon whether achieving them leads to *eudaimonia*. No action is morally right or wrong by definition. Right actions maximize our chances of accomplishing right ends. Therefore, before we can discuss actions (Section 3), we

must consider what ends lead to happiness, for those are the ends that are both morally right and should serve as a guide to life. Discussion of ends will consist of two parts. First, we need to look to happiness-in-general and what ends are included or excluded from that set. Second, we will reflect on the selection and modification of ends by individuals.

1.1 Happiness-in-general

For the writing of our manual, we begin at the end; that is to say with our ends. What codes of conduct and strategies for behavior we arrive at will entirely rely upon our reasons for action, the goals at which we aim. The natural place to begin, given Grice's framework, would be with happiness-in-general. As you recall from the previous chapter, happiness-in-general is the set of all ends desirable for their own sakes (I-desirables) that satisfy some open feature F of this set; we can call this set **H**. Not all ends that are desirable for their own sakes are *ex hypothesi* desirable for the sake of happiness.[1] Alas however that Grice offered no suggestions as to what the open feature defining **H**-membership should be. Further, there seems little reason to suspect that we can conjure from the ether either the complete list of membership in this set or the definitive open feature. So it is beyond the scope of this dissertation to grope about like blind Polyphemus for this open feature, which – much like deceitful Odysseus – is liable to sneak past us. We can leave that for some future, epic work.

These limitations aside, all is not lost. We can at least eliminate a few ends from membership in **H**. We can call these **end exclusion rules**. There is also the possibility for some **end inclusion rules**, which would stipulate some ends necessarily are included in happiness-in-

[1] Recall that the set of I-not-H-desirables ends may be empty, though Grice thought it wasn't. For now, I'll follow Grice on this point, but the theory does not necessitate this view.

general. A possible methodology to derive inclusion rules is to consider some ends that necessarily should be included in anyone's own happiness-for-an-individual set $\mathbf{P_x}$ (which, as you recall is a proper sub-set of **H**, $\mathbf{P_x \subset H}$), and so therefore must be included in **H**. There might also be some necessary prerequisites, ends anyone must realize to be happy. This method comes at happiness-in-general backwards, as it were, since Grice began with happiness-in-general and then proceeded to the more specific happiness-for-an-individual. He appears never to have considered such a methodology, and employing it should be a significant advancement beyond his suggestions.

So, what ends can we rule out as inadmissible to the happiness-in-general set? Grice already ruled out any end that is not desirable for its own sake, that is to say ends that are necessarily instrumental ends. However, this tells us little. If we do not know what ends are final (i.e., I-desirables) we cannot deduce what ends are purely and always instrumental, and never final. At best we would have *prima facie* guesses. Intuitively, suicide is strictly inadmissible as an end within happiness-in-general. Prohibition against suicide seems to be one of the few strict, exceptionless rules. Gricean pirots, in setting their own ends, are working to maximize their likelihood of survival. Hence, Grice effectively eliminated suicide by definition. Besides, what would be the point of being *eudaimon*[2] if one were not around to experience it? Therefore, we can posit an end exclusion rule against suicide. From that rule then, we can derive the **self-preservation inclusion rule:** the end of preserving one's own life must be included in the set of

[2] Throughout the rest of this chapter (and beyond) I will use 'is *eudaimon*' and 'is happy' interchangeably. By saying, e.g., that some end is necessary for being happy, I mean considerably more that might be colloquially expressed by 'happy.' The previous chapter and this explain the notion of happiness and being happy that I mean to convey.

ends that make up one's life plan (i.e., happiness-for-an-individual).[3] Because it should be in everyone's set of personal happiness set, self-preservation is a member of **H**.

Though the exclusion of suicide may be both obvious and uncontroversial, it is nevertheless suggestive of other exclusion rules. Not only must one exist to be happy, but one must experience happiness (though that is admittedly rather vague). In review of Grice's work on happiness, Richard Warner (1986) offers two sound proposal for inclusion rules. First, he suggests that for one to be happy, one must believe one has realized one's ends (or enough of them).[4] Second, one must also enjoy realizing those ends. Hence, it would seem we must include in happiness-in-general the ends of preserving one's capacity for realizing one's ends, of recognizing one has realized one's ends, and of enjoying the realization of those ends. We unify them into the **rational capacity inclusion rule**. Being a Gricean pirot requires a minimum level of rationality. The full measure of that minimum level has not been fully explored.[5] But it will at least include the ability to realize ends, to know whether one has realized them, and to enjoy realizing them. Thus, the preservation of one's rational capacities both is a member of happiness-in-general **H** and should be part of everyone's individual happiness set $\mathbf{P}_x$. Both the self-preservation rule and the rational capacity rule are built-in ends, which a Gricean cannot abandon on pain of irrationality.[6]

---

[3] Though it is necessary for everyone to have this end, I have not ruled out altruism. The end of self-preservation is one of many final ends, some of which can be aimed at helping others. All these ends must be balanced, as I discuss in Section 2.

[4] He separately excludes anyone who has not actually realized his or her ends, but believes falsely that they have been realized.

[5] As a reminder of what Grice does tell us about this level, it requires that one be able to set one's own ends. Hence, it involves a certain degree of awareness of one's self and one's environment, and the capacity for long-term planning.

[6] It is worth emphasizing that the capacity for enjoyment is a *rational* capacity that is just as important (*qua* a built-in end) as our other rational capacities, such as long term planning and self-

It is also worth noting that so far we have not required the fulfillment of these ends without exception, only their inclusion in people's life plans. It remains possible for there to be an extreme situation in which it would be both rational and right for one to sacrifice one's own life or rational capacity in order to satisfy another end, perhaps for the preservation of others.[7]

In *Nicomachean Ethics*, Aristotle considers the question whether a person can be called happy before they are dead. Like Aristotle, I answer that happiness is a property properly attributable to the living. In the current system, the answer is easily provided by the addition of a notion of satisfying enough of one's ends (see section 2) and the **pursuit inclusion rule**: one must have the end of pursuing one life plan. An individual achieves happiness by realizing ends in his or her $\mathbf{P}_x$ set. Some of the ends in that set will be complex, taking anywhere from a few years to a lifetime to realize. Nevertheless, in the meantime, the end established by the pursuit inclusion rule is realized so long as one is working towards those more long-term goals. So, one *can* be counted as happy while working to achieve his or her life plans.[8]

Before proceeding to another methodology, there is one more rule to discuss relating to the matter of the difference between ends and desires. In the previous chapter, I related a concern

---

awareness. This is not a pre-dominantly intellectualized theory of happiness. Enjoyment will also be crucial in the personal selection of ends in Section 1.2.

[7] We might at this point want to also include ends of preserving the lives and rational capacities of others as built-in ends of all rational Gricean pirots, as well as exclude from **H** any ends that inhibit such ends as murder, torture, etc., which destroy someone else's likelihood of achieving happiness. However, at present we cannot make a case for either the inclusion rule or the exclusion rule, at least not without explicitly appealing to Kantian notions of rationality and ethics. Therefore, we omit these rules for now, with the fullest expectation that, given Grice's appreciation of Kant, they will be included in a finished theory at some later date.

[8] This is not a guarantee that this person is happy however. Getting there could be too miserable. It seems at least possible (though highly unlikely) for someone to be miserable one's whole life while striving to achieve some great goal, and then upon achieving it, to realize happiness.

Warner (1986) took up about this difference.[9] He preferred to see them as interchangeable, though others had challenged him on this point. Grice did not address the matter. The challenge is certainly sound. I have the end of grading finals, though I do not desire to do so. It is entirely plausible (though admittedly somewhat odd) for someone to have a maximally final end (which he chose for himself) without desiring it. Therefore, we should posit the **desire inclusion rule**: *ceteris paribus* select only those maximally final (I-desirable) ends that *you* desire to realize. If we can choose our final ends, why would we not choose ends we desire? Whenever someone does desire the realization of her final ends, then we can talk about either of her ends or her desires interchangeably.[10]

Another methodology for developing inclusion rules is to look to folk notions of and empirical research on happiness. This is fairly similar to Mill's (1863) notion of desirability being determined by what is desired: if everyone desires something, it is desirable. Experimental work into the nature of happiness is relatively new, with little research done prior to the 1980s (Myers & Diener, 1995). Methodologies tend to vary depending on the discipline of the researcher. Self-reporting studies (known as subjective well-being studies) are not uncommon among psychologists;[11] experimental philosophers tend to use third-person vignettes to plumb intuitions; economists often incorporate non-psychological, quality-of-life factors. And while there is not

[9] See chapter 3, Section 2, note 17.

[10] The desire rule is not properly an inclusion or exclusion rule. It is a normative principle for any defensible moral psychology, to say nothing of rationality in general. It is closely tied to issues of enjoying the realization of ends, as will be discussed below.

[11] See Diener, Oishi, & Lucas (2002) for a review of this field.

universal agreement in their results,[12] there are some relatively uncontroversial findings. Achieving a theshold level of wealth is typically thought of as necessary for happiness (Diener et al., 1995; Diener & Biswas-Diener, 2002).[13] People reporting themselves to be happy in general tend to have a sense of satisfaction with their employment (Cochran, Antonucci, Adelmann, & Coleman, 1989; Freedman, 1978; Michalos, 1986). External factors, though relevant to happiness, typically have less effect than personality factors. Those saying that they are at least somewhat happy tend to be optimistic (Danber & Brooks, 1989; Seligman, 1991), feel in control of their lives (Campbell, 1981; Larson, 1989), have high self-esteem and believe themselves to be above average in some way (Campbell, 1989), and extroverted (Costa & McCrae, 1980).

The methodology generally employed in these studies relies upon whether people describe themselves as happy; hence the emphasis on *subjective* well-being. They place each individual in the position of final arbiter for determining their own happiness. Admittedly, each person is only happy if they believe they are. Nevertheless, this methodology is not without drawbacks. Not only can people deceive themselves, asking someone whether he or she is happy carries with it substantial normative pressure. People generally assume that they should be happy. Being unhappy is not something one typically wants to admit, to oneself or to others. Thus, asking people whether they are happy may in fact skew results, so that more people reporting themselves to be happy than in fact are. Regardless, this fact demonstrates that happiness is a normative

---

[12] For instance, there is debate over whether people tend to remain at relatively the same level of happiness throughout their lives (Latten, 1989; Stock, Okum Haring, & Witter, 1983; Inglehart, 1990) or whether people become noticeably less happy during their late 30s through their 40s and then become happier again (Blanchflower & Oswald, 2006).

[13] Many of these findings are limited to Western cultures. Whether the same results hold true in other cultures is not always discussed.

notion. Even if most people are not happy (contrary to typical results – see Inglehart, 1990), the fact that most people say they are shows that happiness is a goal people normally accept for themselves.

Rational virtue theory, while being informed by empirical research into happiness, has the opportunity to return the favor. There is a real concern of a culture bias in empirical happiness research. It might be that what Americans or Europeans report as leading to a happy and fulfilling life is not at all similar to the notion of happiness common in other cultures. And if there are widespread differences, then the question of which set of people is happier becomes highly problematic for current empirical methodologies. Rational virtue theory can accommodate such diversity, by saying that the happiness-for-a-person sets for people in the same culture tend to have high degrees of commonality. Americans might tend to include the ends of wealth and work, but there may be other legitimate ways of constructing one's $\mathbf{P_x}$ set, ways found in other cultures, that do not include the pursuit of wealth. So, when many in a society report that a certain end or type of end, such as wealth or family, is necessary for happiness, then that end is a member of the happiness-in-general set **H** and is included in the happiness-for-a-person sets for most of that culture. But it is not therefore required that everyone have that end, especially those in other cultures.[14] This aspect of rational virtue theory demonstrates the wide range of culture diversity that is countenanced in what is nevertheless a moral objectivist view.

In experimental philosophy, Phillips, Misenheimer, & Knobe (forthcoming) have a new study on happiness. They presented four different vignettes to four different groups and then

[14] Culture can impose strong conventional norms on people to adopt ends similar to others in the same culture.

asked participants whether the protagonist of the vignette was happy.[15] The results Phillips et al. found indicate that participants' willingness to view someone as happy was significantly influenced by whether Maria enjoys her life, assuming that she has met some pre-requisites first, namely meaningful relationships with other people and working towards meaningful goals. So enjoying one's life is a necessary component of a happy life. Simply accomplishing goals on a list and not enjoying doing so does not lead to *eudaimonia*.[16] Though Phillips, Misenheimer, & Knobe's phrasing is open to interpretation, it also suggests that one must believe one's life is happy in order for it be happy. Both these results conform to some of the inclusion rules we discussed above.

This is not the place for a comprehensive review of all empirical research on happiness, followed by in-depth analysis to gain further insight into the ends serving as members of the happiness-in-general set. It is the larger point that is relevant here. Empirical research and rational virtue theory make fruitful dialogue partners. Not only can the empirical research further inform the theory, but the theory can also serve as a means to interpret data and guide future research. We should also stop here with our nascent search for members of the happiness-in-general set

---

[15] Their experiment does appear to have a significant framing effect involved. Maria, the subject of the vignettes, is described as either "caring" and having "meaningful friendships and projects" and a "great family life" or as "vapid" without "real friendships." Nevertheless, it does appear that their results bear out some common intuitions. It may also be that use of the vignette methodology, which is fairly commonplace in experimental philosophy, may not be able to address questions of happiness without some framing effects involved.

[16] Unfortunately, this is not the conclusion that Phillips, Misenheimer, & Knobe reach. Instead, they fallaciously conclude there is a Knobe effect. They believe that participants are primarily evaluating the quality of Maria's life: whether someone has good life or a bad one plays a role in attributions of happiness, but not of unhappiness. Thus, supposedly there is an asymmetry in attributions of happiness versus unhappiness based on the quality of a person's life. The results found are more easily explained by supposing that both having a "good life" and enjoying one's life are necessary conditions for happiness, in the folk view. Hence, either not enjoying one's life or having a "bad life" will not make any difference in attributions of unhappiness; either way you are not happy. And these are precisely the results found in their study.

before we veer too far from our central task of linking moral psychology with language. It is enough that we now have some idea of what ends are included or excluded and how to proceed in the future. However, it is worth noting that **H** is still wide open; a great many ends might still be allowed in. Most significantly, we have made no assumptions that would exclude a great many ends that are typically censured. Though it seems likely that they should be excluded, it is not necessary to do so for this project.

1.2 Selection of Personal Happiness Ends

Thus far we have focused on what ends are required in one's personal happiness set, $\mathbf{P_x}$, and what ends can be included in that set (by virtue of being a happiness-in-general end). But I have said little about how one selects from among all the diverse ends within happiness-in-general to develop one's own life plan. I take it as a strength of the theory that it countenances (and indeed explains) such diversity in people's hope and goals, the great multiplicity of life plans and ways to be happy. Nevertheless, though we all have different systems of final ends,[17] some of them are better than others. Personal systems of final ends are to be judged on how likely they are to allow their owner to achieve happiness, i.e., realize enough of their final ends. Hence, we need rules for evaluating someone's personal happiness set $\mathbf{P_x}$.

There never has been a precise moment in the course of anyone's life when he or she stopped and chose all the ends for his or her entire life. Our goals, ambitions, and desires change

[17] At this point, we are limiting our attention only to personal final ends, those ends that are I-desirable, H-desirable, and selected by a person for inclusion in his or her own $\mathbf{P_x}$. We will address the addition of instrumental ends in Section 2.

and evolve over time, building upon our past successes and failures.[18] For instance, Herzog, Rogers, & Woodsworth (1982) found that satisfaction with social relations and health tend to become more important to people and their evaluations of their own happiness as they age, a not unexpected result. Grice (1991) suggests that the whole, evolving set of ends over the course of a person's life be taken as his or her life plan, a suggestion I will follow.

So we not only need rules for selecting ends initially (if there ever is such a time), but also **selection rules** for updating one's set of ends. I will now sketch a few selection rules, which can then be used in evaluating an individual's happiness-for-an-individual set of ends. By 'rule,' I in general do not mean a requirement not admitting any exception. With the exception of the first rule, they are all *ceteris paribus* guidelines and evaluative principles. I am not here attempting to give a comprehensive list of rules, but a few examples, some of which are needed for later purposes.

The contents of one's $\mathbf{P}_x$ set are not the only, or even first, point of evaluation. Who it is that possess esa given set of final ends is of greater importance. Though a set of final ends may maximize Jill's chances for happiness, that does not mean that an identical set is right for John. The first rule reflects this point, and is the exception as a strict rule. According to the **individualization selection rule**, for anyone's $\mathbf{P}_x$ ends, they must maximize *that* person's chances for satisfying them; that is, those ends must be appropriate for that person. One obvious application of this rule is the prohibition of simply copying someone else's ends. There can be reasons to adopt the same final ends as someone else, but only for particular reasons, some of

[18] Still, supposing one moment of decision can be a useful fiction when evaluating personal systems of ends. But it is nothing more than a fiction, which can be re-interpreted if necessary. Any subsequent discussion of the matter is to be regarded as such.

which I will come to shortly. Just because Jill is happy with her ends doesn't mean that her friend Sarah should therefore adopt the same final ends.

Along with the individualization rule are two others I will briefly touch on: the enjoyment and risk selection rules. We have already discussed the necessity of enjoying realizing one's ends. Let us take enjoyment as a primitive psychological concept for Gricean pirots not requiring further explanation.[19] Enjoyment from realizing ends is not a binary condition, but comes in degrees. So the **enjoyment selection rule** states that one ought *ceteris paribus* to prefer an end whose realization provides more enjoyment. Satisfaction of this condition allows for preference ranking among ends and speaking of ends desired conditionally, given that the desire rule is also satisfied.

When selecting or changing ends, consideration must also be given questions of risk. With few exceptions, the adoption of an end does not guarantee the realization of that end. The odds can be heavily stacked against it. So, one must consider the likelihood of achieving a given end.[20] Using that probability and the degree of enjoyment experience from realizing that end, one can perform an expected utility calculation. The tools of decision theory and game theory then become useful tools to aid one in the adoption of ends. According to the **risk selection rule**, one ought *ceteris paribus* to select ends with the highest product of probability of being realized and enjoyed. Subsumed under this rule is the question whether one is risk averse, risk neutral, or risk

---

[19] Genitorial work on pirots simpler than Gricean pirots should be able to provide an account of enjoyment. Hence, enjoyment is a genitorially justified concept, and we are licensed to use it here.

[20] Culture can play a significant role here. One's culture can make some ends more feasible and others much harder. There may be a wide variety of the types of $\mathbf{H}_i$ sets that are encouraged from one culture to the next.

seeking.[21] This rule automatically rules out ends impossible for one to realize, since they have a probability of zero to achieve. For instance, it would be the height of folly were I to desire to be the emperor of Rome, discover the secrets of alchemy, or to find the largest prime number. These ends are historically, metaphysically, and logically impossible respectively. At this point, whether something actually is logically impossible or one merely believes it to be so amounts to much the same thing.

Perhaps surprisingly, these three rules do a great deal of work, taking care of considerations for which additional rules might be thought necessary. For instance, we do not need a special rule to address matters of culture: the risk rule should, if properly applied, account for all effects of culture on what ends one should select, since one's culture can affect how likely it is for one to realize an end (or how enjoyable it will be). Likewise, there is no need for a replacement rule, for cases when someone needs to replace an end (now impossible to realize) with a new one. We have yet to discuss the means by which we achieve our final ends, but that will come in the next two sections.

2 Guide for life

So far, I have only dealt with maximally final ends. But any complete system of ends for a person must include purely instrumental ends. And it is jointly one's final and instrumental ends that constitute a guide for life, offering direction for long-term plans and immediate actions. This section will focus on sketching rules for how to develop a guide for life and will proceed in two

[21] See Arrow (1965) and Pratt (1964).

sections, addressing first the selection of instrumental ends and second the balancing all of one's ends in a single life plan.

2.1 Instrumental Ends

The notion of an instrumental end is fairly intuitive, but let us propose a definition for clarity. An instrumental end is any end that a person believes could be contributive to the realization of some other, higher-order end. Instrumental ends are subjectively defined. Because we are developing a theory of moral psychology, we need only be concerned with a person's beliefs: one can be mistaken, wrongly believing an instrumental end to be contributive to the realization of another end. This would nevertheless classify as an instrumental end. Likewise, we are limiting instrumental ends to only those ends believed to be conducive: if some other end would in fact be contributive but is not known or believed to be so, then it does not count as an instrumental end.

Regarding how an end can be conducive, the most obvious (but not only) way an instrumental end can be contributive for the realization of a higher-order end is by a cause-effect relationship between the ends. The realizing instrumental end is a cause whose effect is the realization of the higher-order end.[22] However, Grice (2001, pg. 124ff) notes that this is far from the only mode of contributiveness. He mentions two other modes, specificatory and inclusive contributiveness. For specificatory contributiveness, there are some ends that can be realized multiple ways, such as showing patriotism. One can wave a flag as a specific means of realizing the higher-order end of showing patriotism, but that is not the only way.[23] With inclusive

[22] It may be an initiating, sustaining, or contributory cause.

[23] Grice contends there is a third mode contributiveness, though I believe he is mistaken, and only describing a variant of specificatory contributiveness. He suggests that an instrumental end can be

contributiveness, the instrumental end can either be identical to the higher-order end[24] or included as a necessary but insufficient component in a complex higher-order end. For instance, greeting a friend when he comes to my office is a necessary but insufficient condition for the higher-order end of being friendly. Additional things are required of me to meet that goal. Specificatory and inclusive contributiveness are distinct from the causal modes of contributiveness, unless we stretch the notion of causation beyond the bounds of credulity. Any of these modes of contributiveness are sufficient to qualify as contributive.

Note that the definition of an instrumental end does not require that it actually be adopted. Typically, there are many instrumental ends to choose from when we are deciding how to achieve our goals. Hence, we need some **instrumental selection rules**. Luckily, the enjoyment, risk, and individualization selection rules above (or close variants of them) will do much of the work. Considerations of risk now take on a second component. Not only must one consider the likelihood of realizing an instrumental end, but also how likely that realization makes the realization of its higher-order end. The enjoyment selection rule becomes significantly weaker for

---

contributive by including the higher-order end. For example, taking a cruise can be contributive to realizing a desire to see Naples by the cruise including making port there. However, this is only an example of specificatory contributiveness. There are multiple ways of visiting Naples, and taking a cruise that includes Naples on the itinerary is only one specific way. The only difference is that the cruise is not solely a specific way of realizing the end of seeing Naples, since it also satisfies the desire to go to sea and take a vacation (among others). There is nothing in the nature of specificatory contributiveness that stipulates that the instrumental end realize only one end. Grice's example demonstrates this fact. He says that "waving the Union Jack might be a way of showing loyalty to the Crown" (2001, pg. 125). However, it can also simultaneously realize other ends, such as stating that the ship from which it is flying is a vessel of the Royal Navy.

[24] The possibility of an instrumental end and a higher-order end being identical is Grice's suggestion, albeit a questionable one. If an instrumental end and its higher-order end are identical, in what way are they distinct such that one can be rightly called instrumental to the other? Would not the relationship work the other way, so that either end could be called instrumental or the higher-order end? I suggest inclusive contributiveness omit such odd cases.

instrumental ends. It is still the case that *ceteris paribus* one should select instrumental ends whose realization one would find enjoyable. But the *ceteris paribus* matters more now. There can be plenty of reasons for adopting an instrumental end despite it being thoroughly unenjoyable, such as my grading of finals, given the strength of motivation for realizing the related higher-order end. The way in which an instrumental end is contributive to its higher-order end can make a significant difference. If an instrumental end is but one end among a great many necessary and jointly sufficient instrumental ends, then the lack of enjoyment counts for little, especially if most of the other instrumental ends provide enjoyment. However, if an unenjoyable instrumental end is specificatorily contributive to some higher-order end, then that lack of enjoyment is a more serious matter for one's deliberations. Perhaps there is another instrumental end that would be more enjoyable to realize.

Additionally, one can have an instrumental end of finding some higher-order end enjoyable. For instance, one might have the end of going to the opera. Initially this is only an instrumental end that is contributive to the end of social networking with other opera-goers. However, one might then adopt a second instrumental end of finding opera enjoyable. The realization of this second instrumental end can transform the first instrumental end into a final end which one now finds enjoyable for its own sake (and also for its networking benefits). This scenario speaks to the benefit of a harmony of ends, which I will address in Section 2.2.

There are two other instrumental selection rules I wish to discuss. First, I would like to suggest the **general parsimony selection rule**.

> **General parsimony selection rule**: *ceteris paribus* one should select as few instrumental ends as necessary for the realization of a higher-order end.

Basically, it is best not to make your job harder than it needs to be. Second, there is the **higher-order parsimony selection rule**.

> **Higher-order parsimony selection rule**: *ceteris paribus* one should select the set of instrumental ends containing the fewest levels.

There can be instrumental ends that are contributive to a higher-level instrumental end, which in turn is contributive to a final end. And there can be instrumental ends for instrumental ends for instrumental ends, and so on. Intuitively, the idea behind this rule is that *ceteris paribus* the fewest levels required to realize an end, the more likely the realization will be.

2.2 Balance of Ends

In addition to rules about individual ends, we need rules about the composition of a set of ends as a whole. The balance of all our diverse ends is critical to our chances of happiness. Grice (2001) discusses seven features of an individual's system of ends that are conducive to happiness. In the previous chapter, I presented several criticisms of his list. As such, I have distilled his list down to four rules that are still in the same spirit. We can call them the **four balance rules**.

First is the **global possibility rule**.[25] *Prima facie*, it might seem vital not only that individual ends are possible to realize, but also the entire set of ends collectively be possible. Yet, requiring

[25] This rule is akin to Grice's first feature of feasibility, which seems for Grice to be slightly stronger than possibility. I have, however, shifted to possibility due to need for defining feasibility, which Grice left unsatisfied.

that it be possible to realize all of one's ends together is too strong, because we will not require that anyone actually realize all of them. (See discussion of the principle of enough below.) Instead, this rule requires that at least enough of one's ends be collectively possible to realize, so that when they are realized that person has achieved *eudaimonia*. Prudence further suggests that to be on the safe side, it would be better that more than the bare minimum of ends be collectively possible to realize, lest one of those ends becomes impossible to realize and consequently one's chance of happiness be irrevocably lost.

> **Global possibility rule:** *ceteris paribus* at least enough of set one's set of ends should be collectively realizable in order to achieved *eudaimonia*, and beyond the minimum threshold of enough, there more of one's ends that are collectively realizable, the better.

The second rule is the **global autonomy rule**.

> **Global autonomy rule**: *ceteris paribus* the more the realization of system of ends is under one's own control, the more preferable that system of ends is.

The *ceteris paribus* condition is especially significant here. This rule is not an absolute prohibition of any end, the realization of which is not completely under one's own control.[26] It is not an endorsement of Stoicism. Such an interpretation of this rule is liable to prevent a stoic from having very many ends at all, so that even if they were realized, because of the paucity of his accomplishments, it would perhaps be difficult to regard the stoic as happy. Nevertheless, it would

[26] I did not give a general version of the autonomy rule for each final end for this very reason. Such a rule would state that *ceteris paribus* one should adopt final ends that one has complete control over realizationing. Even with the *ceteris paribus*, it would be too strong.

seem the stoics were going in the right direction. Autonomy is generally preferable, but not at the expense of isolation and exclusion from the company of others.[27]

Third is the **harmony of ends rule**.[28]

Harmony of ends rule: *ceteris paribus* one should minimize the degree of competition between one's ends as much as possible.

Some degree of conflict is inevitable, to be expected, and perhaps even advisable. At the very least, there will always be a conflict for time, since that is one of most limited resource in a person's life, and one cannot do everything simultaneously (no matter how hard we try to multi-task). Further, though there could be someone with only one goal in life to the exclusion of caring about anything, such a person is typically be regarded as odd or even pitiful.[29] A rich diversity of goals is desirable. Yet, some degree of diversity of ends is too great; it is antithetical to happiness.

The harmony rule then has two components. First, *ceteris paribus* ends that support one another are preferable to ends that are either support-neutral or competing. Ends support one another just in case the realization of one (or the process thereof) increases the likelihood of

---

[27] Some studies suggest that meaningful relationships with others are crucial for happiness and that extroverts are much more likely to be happy. For example, see McCrae (1980); Diener, Sandvik, Pavot, & Fujita (1992); Emmons & Diener (1986a, 1986b); Headey & Wearing (1992); and Pavot, Diener, & Fujita (1990).

[28] This rule comes from a combination of Grice's third and fifth features, compatibility and supportiveness respectively.

[29] I do not mean to suggest that such a person of necessity cannot achieve happiness. *Perhaps* they can. The main point here is that since such individuals are by far in the minority, our theory of moral psychology would be too much of a stretch from actual humans if we were to require such single-mindedness of everyone. Our theory must embrace the diversity while keeping conflict to a minimum.

realizing the other end. This relationship may be mutual or unidirectional. Non-supportive ends are either in competition or support-neutral. Ends are in competition if the realization of one (or the process thereof) decreases the likelihood of realizing the other (which again can be a mutual relationship or unidirectional). Ends are support-neutral if the realization of either one in no way affects the likelihood of realizing the other end. The second component states that *ceteris paribus* support neutral ends are preferable to competing ends. The degree to which two ends are supportive or competitive can be measured by how much the realization of one increases or decreases the likelihood of realizing the other. The greater the degree of competition, the greater the other considerations must be to justify the adoption of two competing ends. Thus, the ends of doing well in school and having a rich social life are in competition, but not as much as the ends of running marathons and regularly eating rich, gourmet meals. It would take fewer additional considerations to countenance the adoption of the first set (which seems likely permissible) than the second set (which though perhaps not impossible will be very difficult to achieve both).

The fourth balancing rule is the **flexibility rule**. That one's system of ends should be fairly stable has largely already been built in. An end adopted should not be abandoned without good cause. Nevertheless, it is reasonable to assume that people will at some point in their lives need to change their system of ends (for any of a variety of reasons). We have already discussed some rules to govern the modification of one's ends. But flexibility is also a property advisable in one's whole set of ends. So, *ceteris paribus* a flexible system of ends is generally preferable to a more rigid system.

> **Flexibility rule**: *ceteris paribus* one's system of ends should be sufficiently flexible, such that the set can be changed as necessary in order to achieve *eudaimonia*.

Flexibility is the capacity a system of ends possesses to replace, remove, or add ends without significantly decreasing the likelihood of realizing the remaining ends.

Noteworthy are a few items intentionally excluded from the list of balancing rules. Parsimony has been left out. It does not seem that the smaller your system of ends (in terms of number of ends in the set), the better it is, not even *ceteris paribus*. There may be an upper-threshold, beyond which one has too many ends, but such cases will likely be handled by the harmony rule already. Also excluded is Grice's notion of agreeableness. Besides the fact that it was rather unclear (as discussed in the previous chapter), its function of aiding in the selection of two distinct sets of ends has already been accomplished via our other rules, chiefly the enjoyment rules.

Before proceeding to the next section, I need to briefly discuss what I call the **principle of enough**. It is unreasonable to expect people to realize all their final ends and demand them to do so to achieve happiness. Instead, there is some threshold, the reaching of which is sufficient for *eudaimonia* and beyond which one does not become more *eudaimon*. The location of this threshold – that is, what constitutes enough to be happy – is determined by each individual for himself or herself.[30] Supposing that there were two persons with identical sets of ends, one might be more ambitious or find different levels of enjoyment for the realization of those ends, and as such, need to realize more of those ends to consider herself happy. It is necessary to believe oneself happy in order to be so. It is likely that what constitutes enough can change throughout a person's life,

[30] That said, one's level of enough can be culturally influenced. Those surrounded by highly successful and ambitious people will likely feel the need to accomplish more in life to be happy than those who are only exposed to others who accomplished less in life.

typically increasing with age so that was enough for a 30 year-old will not be enough for that same person at 60.[31]

3 Principles for Behavior

Thus far the theory I have been sketching is predominately a sort of virtue theory. The sort of person you are, most importantly the kind of ends you have, are of primary importance. But unlike most types of virtue ethics, this theory also posits normative principles for one's behavior. Having the right sort of ends that aim at happiness is all well and good, but if one does not behave well, he has little chance of ever seeing happiness, and his best intentions be damned.

Kant famously tried to demonstrate one's duty from the categorical imperative by considering what if everyone else had the same maxim as you. If everyone cannot (or you would not will them to), then neither should you. My methodology runs the other way. *Given what everyone already does and is likely to do in the future, what is optimal for you to do?* Then, because everyone is capable of asking the same question (and as rational agents, should ask it), certain

[31] A number of studies (e.g., Blanchflower & Oswald (2006, 2004) and Clark (1996)) suggest that well-being typically is "U-shaped" over the course of a person's life, declining to a minimum in a person's 40's and then increasing again. The theory of moral psychology emerging here can offer an explanation of this phenomenon. It is possible that people's goals become more complicated during their 30s and 40s, and simultaneously the happiness threshold moves too. Then often beginning in their late 40s and moving onward, their system of ends have become more stable and they are realizing more of those ends (for example, promotions, greater financial independence, seeing children grow up and move out), and as such become happier, getting closer to that threshold as they age. Research looking for increasing complexity in people's life plans through their 30s and 40s would be needed to test this hypothesis.

similarities will emerge. First, there will be some general consensus about how to answer this question. These methodological overlaps will go into our manual as principles for deciding on how to act in the world. And even if we come up with different answer, we all get there using the same procedure. Second and more interesting, some general patterns of behavior will emerge as the dominant strategy for everyone, regardless of ends. Some types of behavior just lead to human flourishing for each person separately. These overlapping answers are behavioral principles that we write into the manual.

That comparison is a rough sketch and does not account for all the types of principles for behavior. Before explaining my methodology in greater detail, let me first define behavioral principles and discuss the three varieties of them.

> **Behavioral principle**: a *ceteris paribus* strategy for action capable of repetition and replication, which is stable and rational.

Evolutionary game theory (see Bicchieri, 2006; Skyrms, 1996 & 2004; Harms & Skryms 2008 for example) will provide precise descriptions for many of the terms in this definition. Rational virtue theory advocates making extensive use of evolutionary game theory (EGT) in developing behavioral principles, while offering some advancements as well. The next chapter, where I focus on cooperation, will serve as an example.

3.1 Definition of Behavioral Rules

As with rules about ends, behavioral principles are *ceteris paribus* in nature. Typically, it is in one's best interest to adhere to behavioral principles. Your likelihood of realizing your ends is generally higher if you adhere to behavioral principles than if you do not. For instance, if we were

to claim that telling the truth is a behavioral principle, this would not amount to also claiming that no one can ever get away with lying or realize his ends (immediately or in the long-run) by lying. However, it would follow that lying usually doesn't help one achieve happiness in the long term. Instead, being truthful would consistently be the strategy that increases one's likelihood of being happy. This is just a hypothetical example, but one that indicates an important point. Rational virtue theory does not make any pre-theoretic assumptions about what behavior is right or virtuous. This approach is contrary to Aristotle who began with many assumptions about the nature and moral rectitude of several virtues, from which he derived a theory of happiness. Just as rational theory works in reverse of Kant's method, so too with Aristotle. We begin with a theory of happiness and then work to infer behavioral principles, which are whatever sorts of behavior that *ceteris paribus* enhance one's chances of being happy.

I will now explain each term in the definition of behavioral principles. Repetition and replication are similar, though related to different audiences. Behavioral principles are not unique, one-time situations. They cover a multitude of related scenarios in which one is liable to find oneself. It must also be the case that you are able to repeatedly perform the action called for by a behavioral principle in all (or most) relevant circumstances. Behavioral principles have wide application for there can be a range of similar circumstances in which they are applicable. Requiring that behavioral principles can be replicated means that others, seeing how well following a rule works for one person, are able to replicate the same behavior, adopting the rule for themselves. Replicator dynamics is one of the key tools employed by EGT. (See Skyrms 2004.) Essentially, the idea is that a successful behavioral strategy, one that is stable and increases utility, will spread, being picked up by others. Success breads imitation.

That a behavioral principle should keep our long-term interests in mind should not be surprising. Any strategy that wins the battle but loses the war will not do. Instant gratification at the cost of *eudaimonia* also cannot be countenanced. Hence, it is not enough for a behavioral principle to be a consistently efficient way of realizing instrumental ends. That said, the principles governing the selection of ends discussed above should deal with much of this worry by requiring that our instrumental ends and final ends align. If followed, they should set one's utility payoffs in individual games such that the rational action in that game is not typically counter-productive to our long-term interests. Nevertheless, behavioral principles should follow the same pattern. Behavioral principles can require what appear to be short-term sacrifices for the sake of long-term gain. The main point here is that while the right action and the rational one might seem to diverge for a particular action, they do not from the perspective of a life plan.

To say that a behavioral principle is rational means that performing the action(s) it advocates is in your self-interest, where self-interest is defined by the ends we choose and *eudaimonia* is in the self-interest of us all. Maynard Smith and Price (1973) proposed the idea of an evolutionary stable strategy (ESS).[32] For any game, an ESS is any strategy that, if a sufficiently large portion of the population playing that game play this strategy,[33] then no one playing a new (also called mutant) strategy can invade, i.e., mutant players can do no better than those playing the ESS, and typically will do worse so that no rational player will have reason to switch to the new

---

[32] Since Maynard Smith and Price (1973), there have been various refinements on the notion of ESS, which allow for stability in some additional cases than just those covered by evolutionary stable strategies. Still, ESS is sufficient for conveying the general idea of stability for now. We can leave until the next chapter any more recent refinements we might need for cooperation.

[33] What counts as sufficiently large can be mathematically determined based on the specific utility pay-offs for that game and percentage of the population playing each strategy.

mutant strategy. For example, in the standard Prisoner's Dilemma scenario (see Table 1), there is only one ESS, namely *Defect*.

Table 1: Prisoner's Dilemma

	Cooperate	*Defect*
Cooperate	2, 2	0, 3
Defect	3, 0	1, 1

No one playing *Cooperate* can do better than those defecting. However, if it becomes an iterative version of the Prisoner's Dilemma, where there is a fixed probability of playing the game again, then new strategies emerge. For instance, besides *Always Cooperate*, and *Always Defect*, there is also *Tit-for-Tat*, wherein you first play *Cooperate* and then in each subsequent game you play whatever strategy was just played by your partner (*Defect* or *Cooperate*) in the previous round. If a large majority initially play *Tit-for-Tat*, then that population cannot be invaded by mutants playing *Always Defect*, and therefore *Tit-for-Tat* is an ESS under these conditions. Interestingly however, if the majority play *Always Defect*, mutants playing *Tit-for-Tat* cannot invade and are outperformed by the permanent defectors. So stability is partially determined by what everyone else is doing initially. Thus, the stability requirement eliminates some possible behavioral rules. For instance, suppose everyone (or even a sizeable minority) found it H-desirable to be a philosophy professor. If everyone pursued that career, then doing likewise would not be a stable strategy, since most people would fail to achieve their goal and would have been better off playing different strategies.

3.2 Types of Behavioral Principles

Before attempting to discover behavioral principles, we need a methodology for discovering behavioral principles. To that end, let us divide and conquer. There are two varieties of behavioral principles, based upon how they differ with end selection rules. First are behavioral rules that are adaptations of end selection rules. For instance, we can posit the **behavioral principle of autonomy.**

> **Behavioral principle of autonomy**: *ceteris paribus* act in such a way as to increase your autonomy.

Our reliance upon factors outside of our control can be limited not only by what ends we select for ourselves, but also how we act. A student's failure to study for an exam is an act that has decreased his autonomy, leaving him reliant upon the mercy of the professor for a passing grade, whereas, had he studied, he would have increased his autonomy.[34]

The second type of behavioral principle (and the majority of them) is the best means for realizing the ends previously selected. They are to help us make decisions about how to act, given what ends we already want. Bicchieri (2006) finds two routes to making a decision about how to behave, the "deliberation route" and the "heuristic route" (pg. 4-5). The deliberation route is the method recommended by traditional rational choice theory. It requires a colossal amount of information for every decision. One must investigate all possible outcomes, the probability of each occurring, the payoffs to everyone involved, and then calculate the expected utility of each option. The right choice is whichever one maximizes your utility in that game. Aside from the fact that gathering all that information, if even possible, is very costly in time and other resources,

[34] Of course, this is only a *ceteris paribus* rule. Full autonomy is neither attainable nor advisable; there are times when relying on others is necessary and occasionally even beneficial.

behavioral game theory (see Camerer, 2003) shows people regularly and systematically do not follow the models of traditional rational choice theory.[35] People are not complex difference engines capable of continuously making elaborate calculations to determine the optimal action at every turn. Any theory of rationality or moral psychology that requires them to act as if they were is doomed to failure.

Alternatively, Bicchieri proposes the heuristic route to decision-making, according to which how we typically behave is based upon behavioral habits and social norms. Based on testing by Nosofsky & Palmeri (1997), Lamberts (1997) suggests that in a given situation, we typically look for a similar situation and behave in way that we take to be "appropriate" for that type of situation. The idea is that we have default behavioral responses for types of situations. It is essential then that one have appropriate default behaviors, a helpful notion of similarity or types of situations, and ingrained habits to act upon those default behaviors. The manual details what the appropriate default behaviors should be. Practical wisdom should provide an idea of how to correctly compare situations. By making habits of adhering to the behavioral principles, we will usually behave in ways that are typically conducive to realizing our ends.[36]

---

[35] Camerer (2003) is the standard for behavioral game theory. He presents this data not to dismiss traditional game theory, but to improve it.

[36] My emphasis on habituation is contrary to a common theme in contemporary virtue ethics (cf., Annas, 2002), which downplays the role of habit, stressing the importance of deliberation in spite of habituation, giving these theories a distinctly Kantian tone. This view seems wrong-headed. When I tell the truth to an innocuous question, such as asking for the time, I do so automatically without deliberation and out of habit. Prevarication, however, typically involves forethought. Annas thinks deliberation helps defend against situationism. However, it seems to largely concede the argument to the situationist challenge. An increased role of habituation should offer additional lines of response.

Of course, just as people are not difference engines, neither are we thoughtless automata, mindlessly following behavioral defaults. Biccheiri rightly concludes that the correct decision-making route is a combination of the deliberation and heuristic routes, which she terms rational deliberation. Behavior is not always without deliberation. Some situations are sufficiently unique, either because they are radically different from anything else or they involve an uncommon bit of knowledge or other factors. At times we should pause and deliberate (and a person of practical wisdom should know when that would be). And in some of those cases, all things considered, it is rational to act contrary to a behavioral principle, which is precisely why they are *ceteris paribus* in nature.

So from where do we get these behavioral principles for us to build habits out of them? Some might require deduction, given ends countenanced by the end selection rules. Others are contingent upon the behavior of others: if enough people regularly follow a strategy in situations of a certain type, then that strategy can become stable and it is rational for you to do the same. Conventional norms are powerful things. Finally, some behavioral rules are simply a matter of human flourishing, best adhered to regardless of whether others do so as well. For instance, even if everyone else consumed nothing but pizza and beer, it would still be best for me to have a balanced and nutritious diet.

Particularly interesting is the second source of behavioral principles. Evolutionary game theorists, such as Bicchieri and Skyrms, work on showing the stability of various existing social norms, such as cooperation or altruism. As Bicchieri points out, these social norms can serve as setting our default behavior that become habituated responses. As she defines them (Bicchieri 2006, pg. 12), social norms exist when a sufficiently large subset of a population know of the

norm, expect a sufficient subset of the population to conform to the norm,[37] believe that they are generally expected to conform, might believe it is sufficiently probable that non-conformers are punished or conformers rewarded, and then do conform on the basis of these beliefs and expectations. Not all social norms require punishment and reward by the society to be stable norms, but the presence of such factor can be crucial for the sustainment of other norms, for social norms can often be contrary to our self-interest (narrowly construed). In Prisoner's Dilemma cases, even if there is a norm is to cooperate and there is no punishment for defecting, then conformity is not rational, and we cannot include it as a behavioral principle. However, Bicchieri explains (2006, pg. 26ff.) how the presence of punishments and rewards can induce conditional conformity: while some will cooperate regardless, conditional norm-followers will only do so because enough others do it for fear of sanctions. These sanctions for defecting and bonus for cooperating transform the payouts for the game, converting it into a standard cooperation game. Such stable norms are then included as behavioral principles.[38]

Brian Skryms (1996, 2004) has focused on the stability of various norms and the evolution of those norms. For instance, consider a two-player Divide-The-Dollar game. There is a sum of one dollar to be divided among two players, and each player simultaneously announces a

---

[37] As Bicchieri points out, this is first and foremost an empirical expectation, since the population has witnessed most people conform to this norm repeated for some time. She states the expectation can also become normative, dividing the notion empirical and normative expectation. Ration virtue theory does not posit such a division. If a social norm is stable, then the normative expectation generally follows empirical. Given that most people do conform and conforming is *ceteris paribus* rational, one ought to conform.

[38] Some social norms might be fairly consistent between societies (cooperation is a likely candidate) and thus are included in IMMANUEL. But many will differ from one culture to the next, and as such cannot be in the manual for everyone. We can then state that there is a general behavioral rule advocating *ceteris paribus* conformity to stable social norms (whatever they might be). Then in composing sub-manuals for specific cultures, we can include specific social norms.

percentage of that dollar that he wishes to take. If the total percentage of both players is less than or equal to one hundred percent, then each player receives the percentage he announced; if the total is greater than one hundred percent, both players receive nothing. Intuitively the best strategy is to choose 50%. EGT shows why sharing is the stable strategy for such games, which then becomes a social norm, and should be a behavioral principle.

Imagine a population of 10,000 people. We can simulate this population in a 100x100 grid, where everyone has eight neighbors. Let us randomly distribute throughout the population an initial strategy for the Divide-the-Dollar game of asking for 40¢, 50¢, or 60¢. Each round consists of each individual playing his or her strategy against each neighbor (so playing eight times per round). The game is iterative, and then for each successive game, an individual imitates whatever strategy was the most successful among himself and his neighbors. So if someone (*P*) began with the 60¢ strategy and won three times on the first round, *P* would have $1.80. But if *P* had a neighbor Q, who began with the 50¢ strategy, and Q won seven times, Q would have $3.50 after the first round. Thus, in the next round, *P* would adopt Q's strategy of 50¢, as would any other neighbor of Q that did not do as well as Q.[39] Assuming Q had no neighbor that made more than $3.50 in the first round, Q would stick with his 50¢ strategy. Alexander and Skyrms (1999) show that the 50¢ strategy went to fixation (became the only strategy in the population) in 99.5% of random distributions of initial strategies.[40] Thus they conclude, "Justice is contagious." These

[39] This is a fairly standard use of population matrixes and replicator dynamics, both widely used tools in evolutionary game theory.

[40] The only exceptions were for a few "rigged" distributions with sixteen or fewer (out of 10,000) initial 50¢ players, each of which were possible though unlikely random distributions. In such models, an individual is defined by his location in the matrix. How large his "neighborhood" is can vary. However, to my knowledge there does not yet exist a mathematical model that allows for

results provide good evidence that in Divide-the-Dollar type situations, the appropriate behavioral rule is opting for an even split.

4 Conclusion and Responses to Future Criticism

In this chapter, I have developed the outline of a new theory of moral psychology. Though following the spirit of Grice's suggestions, this sketch of rational virtue theory represents a considerable advancement beyond his suggestions. We now have some idea of the contents of the manual for rational human behavior, and we have a good idea about how to continue filling it in. While considerable work remains to be done to complete this theory, we have sufficiently developed the view to move forward. The main goal of this project is to derive cooperation in language from a Gricean moral psychology, and we can set aside the unfinished business and move on to establishing this connection between ethics and language. That will be the project of the next chapter. Before proceeding, however, I will conclude with a final remark about the manual, the nature of rational virtue theory, and responses to two objections that are likely to be raised against rational virtue theory.

4.1 Pragmatic Proviso

---

individuals to move within the matrix into a more desirable neighborhood. This can be important since in some scenarios, fixation does not occur, but pockets of mutant behavior stabilize. One behavioral strategy has come to dominate much of the matrix and is stable; it cannot be invaded. But these mutant pockets, though incapable of invasion, continue to exist because the mutant strategy does well when playing against other mutants. In a future work, we could develop a model that allows for an individual stuck within such a mutant pocket to move outside of it to another location in the matrix, one within the main, stable population. This would more closely track reality and could offer interesting new results.

Following in the spirit of Grice's typically pragmatic approach to all philosophical problems, I wish to inject a bit of pragmatism into our view of the manual, lest we take too Kantian an interpretation of the rules it contains. To be certain, the rules themselves are generally pragmatic: we adopt them because they work. But therein lies the key to their status. For Kant, the categorical imperative can never change, and neither can the ethical conclusions to which it leads us. A lie is wrong and always will be so; nothing can ever change that. The laws of Kantian ethics are written in stone more solid than the Ten Commandments. The same cannot be said for the rational virtue theory manual. The manual is a pragmatic document. We use it (or should) because it works; it improves our odds of happiness. And the rules and principles within the manual are simply the best strategies we have come up with so far. They do not rest upon any metaphysically firm foundation (if there could be such a thing), immune from change or error. It is entirely possible that some new behavioral strategies can emerge that would be better, more trustworthy means to happiness. In light of such new strategies, the old rules will not work anymore. Then, if a behavior principle stops working, we can re-write the manual. My suspicion, however, is that there are some principles (but not all) that work pretty damn well for a reason, and we are not likely to find better ones. Cooperation, I think, is liable to be such a principle.

4.2 Virtue Ethics

I want to briefly discuss why rational virtue theory counts as a type of virtue ethics. First and foremost, rational virtue theory follows in the Aristotlean and neo- Aristotlean tradition of emphasizing *eudaimonia*. It is the driving force of the entire theory. Second, emphasis is placed more on what sort of person you are than what you do. And what sort of person you are is defined first by the ends at which you aim.

Secondary and subservient to those ends are the behavioral principles. These behavioral principles are not strict deontic rules. They are general guidelines for how to behave, or more precisely what sort of person to be in general. In the next chapter I will argue that cooperation is a legitimate behavioral principle. That does not mean that one is obligated to cooperate all (or even most) of the time. Rather, it means that one is cooperative at the right time and in the right way will be better off, i.e., more likely to achieve *eudaimonia*. So these behavioral principles are virtues, or at least virtue-esque.

4.3 Response to Potential Objections

There are at least two objections that could be raised against rational virtue theory, egoism and situationism. The former questions why one should adhere to the manual; the latter suggests that we cannot. Without dealing with either in depth, I will offer a few brief remarks to each, foretelling fuller responses that can follow later.

The challenge of the egoist is one Grice foresaw, relating his thoughts to Grandy and Warner (1986, pg 34ff). We can imagine an egoist who questions why he should conform to the behavioral principles laid out in the manual, particularly if they ever require him to behave altruistically, contrary to his own, immediate self-interest. As should be clear from the previous discussion, this egoist's objection would be based upon a misunderstanding of the demands of morality. Rational virtue theory is a form of egoism. The principles it advocates are based upon self-interest and are the most reliable means for achieving the self-interested aim of happiness. Anyone choosing to forsake the demands of morality does so on pain of irrationality.

Another objection comes from situationism. Based on experimental results from social psychology, situationists (cf., Harman 2000, 2001, Doris 2002, and Vranas 2004) are skeptical of character. They claim that people lack stable and robust character traits. Instead people's behavior can be radically influenced by non-morally relevant situational factors, such as finding a dime. Therefore, if people do not have traits, then they cannot have virtues. Therefore, virtue ethics must be wrong. The situationism debate is widespread and ongoing at present, and I cannot here offer a full response. Some remaining proponents of virtue ethics (cf. Russell 2009 and Snow 2009) have mounted powerful defenses, which rational virtue theory could build upon later. For now, I will only point to a few replies to situationism that rational virtue theory opens up.

First, an underexplored response to situationism is moral elitism, which claims that the majority of people are not especially moral. Hence, it is not surprising that social psychologists have not found widespread evidence for character traits and virtue throughout society. Though present in Aristotle, moral elitism is anathema to much of contemporary virtue ethics, which emphasizes (or tacitly assumes) a moral egalitarianism, where most everyone, if not already generally good people, at least possess the capacity to become morally good.[41] Yet, rational virtue theory's linking of morality (and character) with rationality provides further credence to moral elitism. We would not be surprised by experimental results showing people to be irrational. We should expect such results. If behaving morally is behaving rationally, and most people regularly

[41] This response to situationism is not unique rational virtue theory, and in fact rational virtue theory will have to work to incorporate it, given my views in the next chapter about widespread cooperation in language. Still it can be incorporated by claiming that people have a greater tendency to be cooperative in communication, but not in many other situations. This would be a limited moral elitism.

behave irrationally, then most people regularly behave immorally. Therefore, we should not expect widespread evidence of character.

The second line of response opened by rational virtue theory takes a different tack. In the next chapter, I will argue that cooperation, central and necessary to communication, derives from our moral psychology, such that cooperative communicative behavior is a particular form of moral behavior. The ability to meaningfully communicate is ubiquitous in human societies. If Grice and I are correct, then engaging in meaningful communication requires cooperation. Therefore, cooperation should be a widespread character trait. My theory thus predicts that cooperation is a widespread and robust trait largely immune to situational framing effects, contrary to traits previously tested for, such as helpfulness. Future experimental work can be done to test these predictions as a reply to situationism.

A third response is that virtue ethics (or at least rational virtue theory) does not require fully robust character traits immune from the effects of non-moral factors. We all have our limits, even moral saints or sages. But a virtuous person limits the extent to which these non-moral stimuli influence his or her behavior. This can mean decreasing the degree or the frequency to which they affect him or her. For instance, when exposed to loud noise she may still become rude with others, but only slightly. Alternatively, maybe it is only occasionally that an annoying, non-moral stimulus prompts a virtuous person to not be helpful, instead of typically not helping as most people do.[42]

[42] This last point on frequency speaks to the need for longitudinal situationist studies. Typically, the experiments involve taking a group of people and seeing how many behave virtuously in one particular instance depending on whether or not they are exposed to the non-moral stimulus.

4.3 Conclusion

Rational virtue theory is far from complete. But we do have a good idea about how to proceed. We have found some principles for our moral psychology. And there are methodologies for the discovery of more. A central claim of rational virtue theory is that by definition what is expedient and what is moral right coincide. Cases where the two appear to diverge likely stem from a misconception of expediency or moral rectitude. With this groundwork laid down, we can move on to investigate cooperation as behavioral rule and see if Grice's conversational maxims can largely be derived from a Gricean moral psychology.

More interesting would be to follow a group of people over time to see if any of them regularly behaved virtuously despite the presence of known non-moral triggers (such as loud noise or un-related rewards). My prediction would be that most people would not demonstrate regular and robust virtuous behavior (and some regularly un-virtuous). However, if we found a few that did, that would be a strong point in favor of moral elitism. It would indicate that it isn't the case that we are incapable of possessing robust character traits, but rather that most of us just lack the traits, not the capacity for them.

Chapter 5: Rational Cooperation in Language

So far I've developed a theory of moral psychology, but not yet established its connection with language. To do that, we need to show that cooperation is a behavioral principle, i.e., that it is a stable and self-interestedly rational strategy for action. In this chapter, I will demonstrate how cooperation in the use of language use can be derived from Rational Virtue Theory, thereby establishing linguistic cooperation as a behavior principle in our moral manual. I will begin with cooperation in general, reviewing the work of Michael Tomesallo, Brian Skyrms, and Cristina Bicchieri. Then turning to cooperation in language use specifically, I will review Grice's Cooperative Principle, arguing that as stated it cannot work. I present two distinct forms of cooperation in language use and show that at least one of them is a stable and rational behavior principle one ought *ceteris paribus* to follow. Finally, I will demonstrate that the conversational maxims derive from a notion of cooperation different from the one Grice assumed.

1 Cooperation in General

Cooperation can be taken in two ways: coordination or altruism. The former involves working with someone else because it is in your self-interest; the latter involves behaving against your own self-interest for the sake of another. Genuine cases of coordination, when they arise, are rational. It is the most efficient means for each of us to get what we individually want. The only key is properly recognizing coordination cases every time (or almost) and developing a behavioral principle that becomes (typically) operable at the right time. But why should we ever be altruistic? Why do people behave altruistically? Purely altruistic behavior, by definition, leaves one worse off for helping someone else.

I will briefly review some of the literature on cooperation, which generally tries to justify cooperation as a (typically) rational or right thing to do.

1.1 Evolutionary Justification of Cooperative Behavior

Michael Tomasello (2009) investigates the evolutionary origins of human cooperative behavior. His research is primarily based upon comparison of behavior between chimpanzees and human children, typically between twelve and twenty-eight months old. By analyzing the differences in their behavior in similar situations, Tomasello works to derive an evolutionary explanation for why humans cooperate. His research indicates that cooperation emerges in children's behavior initially instead of being an acculturated trait. However, as children grow older and become aware of social norms and values, the manner and circumstances in which children cooperate later becomes culturally mediated.

Tomasello begins by distinguishing three separate types of altruistic cooperation, which appear to involve separate processes and have different evolutionary histories. The first he calls Helping, where some (typically small) amount of energy is sacrificed for the benefit of another. Of twenty-four human eighteen month olds, for example, twenty-two of them helped a newly met, unrelated adult with a trivial problem, such as opening a cabinet door or fetching an out-of-reach object, regardless of any reward (2009, pg 6-8). This behavior, he notes, requires that the child first recognize the adult's goals (something we'll see to be important for implicature and successful communication). Chimpanzees often demonstrate similar behavior, helping in at least some of the same cases, like fetching an out-of-reach object. Cooperation-as-helping, therefore, would seem to be an evolutionarily old behavior that pre-dates humans (though human beings have expanded on the behavior).

Sharing and Informativeness – Tomasello's second and third types of cooperation – are different from Helping since toddlers engage in this behavior while chimpanzees do not. Tomasello notes studies in his lab where pre-linguistic children are cooperatively informative by pointing out hidden objects that an adult is seeking. Chimpanzees, on the other hand, regularly fail to be similarly informative to researchers or other chimpanzees.[1] Sharing, unlike Helping, involves a more tangible resource (such as food) to aid another. Chimpanzees also do not exhibit sharing behavior, particularly when it comes to food, while children in his studies generally do share. Tomasello is convinced that there must be an evolutionary explanation for why humans

[1] Tomasello notes that chimpanzees will sometimes point when it is an object they want for themselves, but not when it is an object sought by the researcher for his own purposes. He argues that this pointing is then best understood as an imperative rather than being informative.

developed these cooperative behavioral tendencies (while other primates didn't), which helped our early human ancestors survive.

Particularly interesting is Tomasello's discussion of cooperation as coordination. He notes that though chimpanzees are social hunters, they do not act in coordination or with a joint purpose. In the Tai Forest of Côte d'Ivoire, an initial chimpanzee will begin a hunt by driving red colobus monkeys. Others will then will join in, some to chase the monkeys, some to block escape routes, and finally an ambusher to complete the hunt. Tomasello argues that this isn't true cooperative behavior however. As evidence, he cites Warneken and Tomasello (2006), where chimpanzees were given four collaborative tasks (two social games and two problem-solving tasks with concrete goals) with a human researcher, who after some time was to cease participating. Consistently, the chimpanzees would not engage in the social games at all. They would coordinate with the researcher on problem-solving tasks, but only until the researcher stopped participating. The chimpanzees then stopped as well and made no attempt to recall the researcher to the task. When they performed the same test on eighteen to twenty-four month-olds, the differences were remarkable. Not only did the children participate in all four tasks, they would prompt the adult researcher to resume the activity, indicating a joint goal. The children also regularly initiated the task again, indicating that they'd transformed it into a social game.

The wealth of data that Tomasello presents prompts him to conclude that at some point, humans must have been subject to some selective pressure to collaborate – likely for food – in a way that our closest primate relatives never experienced. He notes that the white parts of *homo sapiens* eyes are three times larger than other close primates, a factor key in being able to follow someone's gaze and coordinating on gathering of food and avoiding danger. Human children

quickly become adept at following the eyes of other humans, while chimpanzees do not. Tomasello hypothesizes that the fact that one's eyes can be followed could not have evolved in an environment in which it was likely to be exploited by others, but instead lends itself to and is only possible in a cooperative social environment.

The key point for our present purposes is that cooperation in a variety of forms is a human behavior that is widespread and common in human beings from a very early, even pre-linguistic, age. The way in which we cooperate will become moderated and informed by social norms (a point we will consider more below), but for some reason or another cooperation is part of human psychology. To answer the question why, Tomasello in part turns to the work of Brian Skyrms, an example we will follow.

1.2 Game-theoretic Justification of Cooperative Behavior

One way to investigate the evolutionary origins of human cooperative behavior is to model situations in which such behavior is both stable and increases one's chances of survival and procreation. This is precisely what evolutionary game theory works to accomplish. Brian Skyrms (2004) develops an argument to demonstrate the reason behind cooperative behavior.

Skyrms begins by considering the Prisoner's Dilemma.

Table 1: Prisoner's Dilemma

	Cooperate	*Defect*
Cooperate	2, 2	0, 3
Defect	3, 0	1, 1

In a Prisoner's Dilemma scenario, each player (*Column and Row* we'll call them) can either *Cooperate* or *Defect*. In *Table 1*, the payouts for each of the four possible outcomes is listed, one in each box; *Row*'s payout first, then *Column*'s. As you can see, if they both play *Cooperate*, they each receive a reward of 2. This option is better for both than the double *Defect* result, where they both only receive 1. So the *Cooperate/Cooperate* option is weakly Pareto optimal, which is defined as a solution concept in a game of two or more players in which there is no other solution concept where all players would receive a higher expected payout (Myerson 1991, pg. 97). Our players, however, are self-interested. If *Row* expects *Column* to cooperate, he realizes he would do even better by playing *Defect*. The rational course of action then is for *Row* to *Defect*. However, *Column* realizes the same thing. So each ends up playing *Defect*. Any time you have a result where neither player has a reason to change unilaterally to a different strategy, because he will end up worse for it, then that result is a Nash equilibrium, which is defined as a solution concept in a game of two or more players in which no player can achieve a higher expected payoff by unilaterally changes strategies (Myerson 1991, pg. 91-98). In this case, *Defect/Defect* is the only Nash equilibrium. For each player, *Defect* is the rational strategy, even though it is not optimal.[2]

Things change when the game is likely to be repeated, as Skyrms (2004) notes. A player's previous actions influence the other player's decisions in subsequent rounds. Reputation matters. Let's assume there are only two strategies: *Always Defect* and *Tit-for-Tat*. If there is a fixed .6 probability of playing a future round, the pay-offs for the game morph into something other than a Prisoner's Dilemma; it becomes a Stag Hunt.

---

[2] For more on the Prisoner's Dilemma, Nash equilibria, and Pareto optimality, see Myerson (1991, pg. 98ff).

Table 2: Changes to the Prisoner's Dilemma with a .6 Probability of Iteration[3]

	Tit-for-Tat	*Always Defect*
Tit-for-Tat	5, 5	1.5, 4.5
Always Defect	4.5, 1.5	2.5, 2.5

Now there are two Nash equilibria. But are they both evolutionarily stable? Will a few *Always Defect* players be able to consistently beat a population full of *Tit-for-Tat* players?

We can answer this question using replicator dynamics, as explained in the previous chapter. But first let's simplify our game. Suppose you can either hunt hare or stags. Hares are easy to hunt and can be caught single-handed. Stags, on the other hand, are more difficult and require a partner, but they also provide greater sustenance. The game looks like this:

Table 3: Stag Hunt

	Stag	*Hare*
Stag	4, 4	0, 3
Hare	3, 0	3, 3

If you hunt hare, you always succeed. With stags, you only succeed if you and your partner cooperate in the hunt. So risk is involved. Why cooperate then? You can't lose if you play *Hare* every time.

We begin, as in the Divide-the-Dollar game from before, by constructing a 100x100 grid of 10,000 individuals. We seed the grid with random distribution of either *Stag* or *Hare*, so that

[3] Skyrms (2004, pg. 5).

initially everyone will play whatever was randomly assigned them. Then the game begins. Each person plays all eight of his Moore neighbors (N, S, E, W, NW, NE, SW, SE) once per round and then sees how he and all of his neighbors did. For every player, if any neighbor did better, a player adopts that neighbor's strategy for the next round. Otherwise, a player keeps the same strategy from the previous round. What we are looking for is to see if a stable picture emerges. Let's suppose we started with a 50/50 split between *Stag* and *Hare* randomly distributed. We run the test over and over with different random 50/50 distributions. Fascinatingly enough, 99% of the time *Stag* eventually becomes not just the predominate strategy, but the only strategy that anyone is playing (Skyrms 2004, pg. 32). Cooperation wins out.

There does, however, need to be a sufficient number of cooperative stag hunters to begin the game for cooperation to be stable. For instance, if we have only 10% of the population begin as stag hunters, the end result is almost always a population of only hare hunters. And when we start with 30%-40% playing *Stag*, a stable pattern emerges of rectangles of stag hunters surrounded by hare hunters, with neither group being able to successfully invade the other (Skyrms 2004, pg 37). With the payouts as listed on *Table 3*, there is little difference between the payouts for *Stag* and *Hare*. If that gap widened so that successful stag hunting paid a higher reward, the percentage of initial stag hunters needed would fall as well, since groups of hare hunters would be increasingly susceptible to invasion. What this use of replicator dynamics shows us is the rationality of cooperation. Just look around you. If many others around you are cooperating on a joint goal, it's probably in your self-interest for you to join in too.

Cristina Bicchieri (2006) explores another means by which Prisoner's Dilemma games are transformed into cooperative games. The presence of a social norm can greatly influence our

preferences in Prisoner's Dilemma games. The norm may be as simple as a player's expectation that enough of the rest of a population will follow a behavioral rule and the belief that most others expect him to observe it as well. Or the social norm may come with a sanction for those who do not adhere to the rule (or a reward for those who do conform). A society may regularly and perhaps harshly punish defectors, in which case defecting is no longer in one's self-interest. Bicchieri (1997) gives the rationale for societies to establish such social norms. Using replicator dynamics,[4] she demonstrates that with a high enough probability of punishing defectors and a high enough penalty, cooperation dominates a population, so that defecting in a Prisoner's Dilemma game becomes nearly non-existent. And cooperation dominant societies have a higher aggregate wealth (when the payout is understood monetarily) than a defection dominant society, which results from a much lower rate of punishing defectors. So, for a society to establish and enforce a social norm is rational. It is worth noting, however, that a key to Bicchieri's analysis is that in her model the penalty for defection comes not from players within the population, but an *external* referee who assesses fines. That said, the main point about the transformative role of social norms remains.

The upshot of all this is that cooperative behavior in general looks to be genuinely in one's self-interest in many scenarios. We should first look around our society to observe what others are doing and whether there are any social norms or sanctions at work. Doing so, we find a great many occasions for cooperation. It really does behoove us to cooperate with traffic laws, social etiquette, and respect for private property. But cooperation need not always be warranted. If by

[4] Bicchieri admits that her version of replicator dynamics is more Lamarkian than Skyrms's Darwinian model.

some cosmological fluke, you found yourself transported to a barbarous and anarchic time following the same cooperative norms and expecting others to do the same would appear pure folly. Thankfully however, so long as you don't find yourself among barbarians, behaving cooperatively is generally a rational behavioral strategy, all the more so since we do not enjoy the luxury of constant calculations and comparisons to our neighbors as players in Skyrms's evolutionary games must do.

2 Cooperation in Language Use

When it comes to language use, cooperation is a more complicated affair. We have already given a brief justification of cooperation in general. But to justify as rational cooperation in language use, the same lines of reasoning will not work, as I will soon explain. A new approach is needed. To that end, I suggest we turn to Grice who points us in the right direction.

Grice (1975a) contends that conversation should be understood as a rational activity, so that when we engage in conversation we have a particular purpose for doing so. For it to be a conversation, we must have some dialogue partner(s) as well. That person will, for the same reason, have a particular purpose for engaging in the conversation with us. Grice notes that conversations are not "a succession of disconnected remarks" (1975a, pg. 45). Because I am rational and have a purpose for the conversation, there is an order and flow to the series of my utterances. Likewise with my dialogue partner, I can assume his remarks are connected by his purpose. Since we are speaking to one another (and not just to ourselves), there is a general flow to our conversation; our remarks are connected and ordered. Grice concludes then that there must be at least to some extent an overlapping purpose between us. The fact that we share some

accepted purpose or direction is what makes conversation possible. Cooperation is a necessary condition. Such considerations lead Grice to postulate his Cooperative Principle: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (1975a, pg. 45). Deriving from the Cooperative Principle are conversational maxims: Quality, Quantity, Relation and Manner. In general, Grice claims, observing the maxims and their submaxims (e.g., "Make your contribution as informative as is required") results in satisfying the Cooperative Principle. And it is in virtue of your being cooperative that others are able to work out what you implicated that you so not *say*.

While I think there is much that is right in Grice's account of conversation and cooperation, for our present purposes there are two significant problems. The first problem is not with the Cooperative Principle itself, but the unlikelihood of being able to justify it as a rational behavioral strategy using evolutionary game theory. Some conversations are cases of coordination, where to achieve a common goal it is in one's best interest to cooperate with other conversational participants (DPs from here on). But this is not always the case. There can be plenty of competitive conversations without a sufficiently common purpose or direction. Therefore, any game-theoretic justification here would have to be bifurcated into conversational coordination and conversational altruism. Intuitively, the former – which would involve a shared goal – would be a rather straightforward analysis.

Conversational altruism, on the other hand, would be a different matter altogether. It entails some self-sacrifice to the benefit of someone else. To simplify matters, we can assume that the beneficiary of your altruistic sacrifice to be one's DP. So, why should you cooperate? A fixed

probability of repetition and use of Moore neighbors doesn't seem applicable to language use in the real world. It is not uncommon to have a conversation with someone with whom there is little to no chance of ever seeing or speaking to again. Using external sanctions to transform mixed-motive games into coordination games (as Bicchieri does) also won't work for language use. Here, however, we have no such external, sanctioning referee. Even if there are social norms for cooperation in conversation, it seems that any sanctions imposed on non-cooperative players are too infrequent or weak to be potent enough to transform mixed-motive conversations into coordinated ones. Even then, the sanctions are internal, meted out by the players themselves instead of by an external referee as Bicchieri describes. Other models could be produced which rely on internal sanctions of non-cooperative behavior by the players themselves. Nevertheless, the size of the sanction is likely to be greater than what we find in the real world. A quizzical look by one's DP (or even a stern scolding) would not necessarily be sufficient to induce you to cooperate if it wasn't already in your interest to do so. The penalty must be sizeable and frequent enough to make non-cooperation costly.

Conversations come in a wide variety of formats. Some are distinctly coordination games, with cooperation being the only Nash equilibrium. Other times we face Stag Hunt type conversations: cooperation leaves everyone better off, but defection by everyone can work as well. Then sometimes there are Prisoner's Dilemma conversations, where cooperation is distinctly disadvantageous. Politicians face this from the press: they'd sometimes (often?) rather not be forthright and forthcoming in their answers to reporters (Plüss, 2010). The journalist and the politician each have distinct purposes for the conversation. And yet the tradition of interviewing politicians survives.

So we might conclude that the best way to proceed would be to develop a taxonomy of the types of conversations vis-à-vis cooperation and then develop game-theoretic explanations for why cooperation is (or isn't) rational in each.[5] While this might be an interesting exercise, let's first see if we can still say something more universal and expansive about conversations in general in the face of different conversational goals. This leads us to our second problem with Grice's analysis, which centers on this question: can you non-cooperatively conversationally implicate something? Or rather, does your observing the conversational maxims equate to observing the Cooperative Principle?[6] A non-cooperative conversational implicature is a new notion. If possible, it would mean that a speaker could refuse to cooperate with the purpose or direction of a talk exchange (typically as set by his or her DP), but nevertheless still be able to generate a clear conversational implicature. And that non-cooperative conversational implicature would be able to be explained in terms of the conversational maxims and on the assumption that the speaker was observing or flouting (obviously violating) them. The question boils down to what is the relationship between the Cooperative Principle and the conversational maxims. Are the maxims derived from the Cooperative Principle? Is observing the maxims a sufficient condition for observing the Cooperative Principle?

---

[5] One might object at this point by claiming that for Grice the issue is really about assumed cooperation. Indeed, speakers do have to assume their DPs are being cooperative, as Grice makes clear and I discuss in the next chapter. However, for that assumption to be warranted, it must be the case that speakers actually do cooperate sufficiently often for that assumption to be worthwhile. If conversations are regularly Prisoner's Dilemma cases (where cooperation is not in one's rational self-interest), then speakers either won't cooperate often enough or they are irrational fools for doing so. The cooperative assumption needs some foundation on which to stand.

[6] As Grice notes, you can observe the Cooperative Principle without observing (all of) the conversational maxims, such as in cases of flouting a maxim or clashing maxims. (See chapter 6 for more on this.) But the question is whether someone observing all the maxims is thereby necessarily observing the Cooperative Principle.

Grice has little to say about this relationship. He connects them together in one sentence, stating "On the assumption that some such general principle as [the Cooperative Principle] is acceptable, one may perhaps distinguish four categories under one or another of which will fall certain more specific maxims and submaxims, the following of which will, in general, yield results in accordance with the Cooperative Principle" (1975a, pg. 45). This suggests that he believes observation of the maxims is sufficient condition for observing the Cooperative Principle.[7] It also suggests that Grice regards the maxims as deriving from the Cooperative Principle. In his "Retrospective Epilogue" Grice reinforces this view when he states that maxims are "dependent" on the Cooperative Principle (1987, pg. 368). That view, however, doesn't hold water.

Cooperation, as Grice conceives of it, is too narrowly defined, and there are countless counter-examples of non-cooperative conversations. They occur all the time in courtrooms, between journalists and politicians, and even among family members. The DPs have radically different agendas, yet they are still able to understand one another and conversationally implicature things. Consider the following case. *A* asks *B* a very embarrassing question. Rather than respond, *B* just looks *A* in the eyes and says nothing. I take it as clear to *A* and anyone watching that *B* is implicating he does not want to answer the question and will not do so. In that case, *B* implicated this without saying a word.[8] *B* has refused to cooperate with the purpose of the

---

[7] I don't wish to become embroiled in Gricean exegesis on this point. His addition of "in general" could indicate that Grice thought it was possible to obey the maxims and still not satisfy the Cooperative Principle. However, he does not discuss the possibility in his (1975a). If he regarded this as are real possibility, the point has in general been missed by his readers.

[8] In a face-to-face conversation such as this, it is highly likely that the meaning is partly conveyed by *B*'s body language, though I've tried to leave that out. To see that the implicature does not depend on the body language, consider the same conversation over instant message, text, or email. If *B* regularly responds very quickly and *A* knows this, then no response after even after a few minutes

conversation as established by *A*'s question. Yet, *B* nevertheless was able to non-cooperatively conversationally implicate his meaning.

If non-cooperative conversational implicature is possible, what then is the relationship between the Cooperative Principle and the maxims? To answer that question, we need to reinterpret what it means to cooperate in our use of language. I wish to propose two different types of cooperation. First there is what I am calling **Conversational Cooperation**, with a corresponding **Conversational Cooperation Principle:**

(ConvCoPrin) **Make your conversational contribution such that it is consistent with the assumption that you are collaborating with your DPs towards achieving a roughly shared or generally overlapping set of tasks or goals.**

You and your fellow conversational participants are working to some roughly shared goal(s), e.g., deciding where to go for dinner that night. Everyone's goals need not be precisely identical – I may have the secondary agenda of saving money that you don't – but the overlap is sufficient that we can conversationally collaborate. Somehow, a purpose or direction to our conversation was set and accepted by everyone, and we both work toward that purpose. The purpose did not have to start out as shared prior to the conversation (or even the current stage of the conversation). You might have a goal in mind – one to which I may be initially ambivalent or hostile – that you induce me to take as my own for the sake of our conversation. Conversational Cooperation is, I take it, how Grice's original Cooperative Principle has generally been regarded. Though Grice did not actually state in his formulation of his Cooperative Principle that the purpose or direction had

---

still conveys the implicature (though, admittedly, *A* can't be quite as certain as in a face-to-face conversation, since there remain possible, though less probable, explanations for the silence).

to be shared, there is reason to think this is what he meant. Certainly, his non-linguistic analogy between the Cooperative Principle and two people working together to mend a car lends credence to this interpretation.

What then of conversations without a shared or similar purpose or goal, i.e., seemingly non-cooperative conversations? The police officer and an accused criminal have radically different goals during an interrogation. One might say that in such a case the accepted and shared purpose of the talk exchange is at a higher-order level – namely to be understood – since that would seem to be the only goal they have in common. Each DP then makes his or her conversational contributions as relevant, truthful, etc. as is required given that purpose. This is entirely the wrong notion. Neither of these DPs, if asked, would claim this as their purpose or goal for the conversation is to be understood. The police interrogator would say she is trying to get incriminating information (or even better, a confession), while the accused is trying to conceal information. (Let's imagine he's guilty.) Does this mean that they can't understand one another's meaning or implicature? Not at all. Suppose she asks "Where were you on the night of the 25th?" He might answer, "Gee, I don't know. The moon, maybe." He has obviously *said* something false. But he has implicated, and expects the interrogator to recognize, that he is implicating, that he is not going to say where he was. And she is perfectly capable of realizing this conversational implicature.

The problem with Conversational Cooperation centers on the notion of an accepted purpose, direction, goal, or task. Whose purpose is it supposed to be? And why do we have to have shared or similar purposes in order to communicate and implicate? The cop and the criminal did not need a shared purpose to be understood. Two people can have a conversation so long as

each recognizes and accepts the purpose, direction, or goal of the other, either beforehand or over the course of the conversation based on their utterances. Whether they share the same goals is irrelevant. With this fact in mind, I propose a different notion of cooperation in language use, namely **Communicative Cooperation**, and with it the **Communicative Cooperation Principle:**

(CommCoPrin) **Make your conversational contribution such that your conversational participant(s) is sufficiently likely to be able to understand your communicative intentions and meaning, and presume that your conversational participant(s) do so as well.**

Contrary to the standard Gricean view, the *main* purpose or goal of each DP is no longer to be understood by the other DPs. They can have just about any goal they want, e.g., to avoid incriminating oneself. A DP is only communicatively cooperative in order to satisfy those goals. Being communicative and clear is not an end, but a means. To support this notion of cooperation, we need to accomplish three things: 1) demonstrate that communicative cooperation is indeed a form of cooperation; 2) justify communicative cooperation based on Rational Virtue Theory; and 3) derive the conversational maxims from communicative cooperation instead of conversational cooperation.

One might think that communicative cooperation is cooperation in name only. How can it be cooperative if all DPs are working with different purposes? They are not working together to achieve anything. The reason why communicative cooperation is cooperation can be seen in considering how Humpty Dumpty wasn't cooperative. In *Through the Looking-Glass*, Carroll records this conversation between Humpty Dumpty and Alice:

"And only *one* for birthday presents, you know. There's glory for you!"

> "I don't know what you mean by 'glory,' " Alice said.
>
> Humpty Dumpty smiled contemptuously. "Of course you don't—till I tell you. I meant 'there's a nice knock-down argument for you!'"
>
> "But 'glory' doesn't mean 'a nice knock-down argument'," Alice objected.
>
> "When *I* use a word," Humpty Dumpty said, in a rather a scornful tone, "it means just what I choose it to mean—neither more nor less."[9]

Humpty Dumpty is not being communicatively cooperative. Alice could not have known that Humpty Dumpty meant 'a nice knock-down argument' when he used 'glory.' So he is a fool if he expected Alice to understand what he meant, and blatantly (and annoyingly) uncooperative if he didn't but said it anyway.[10] Assuming Humpty had any communicative intention, his use of 'glory' in such a non-standard way made it far less likely that his intention could be realized by Alice understanding him. This would hold regardless of Alice's goals and intentions and how they matched Humpty's. So, communicative cooperation does not require DPs to have sufficiently similar communicative goals. Rather, all that is required is for DPs to recognize and respect communicative expectations (such as for word meaning and syntax). A non-linguistic analogy is the formation of coalitions between diverse political parties to form a government. The Conservatives and the Liberal Democrats of Great Britain, who have largely very different aims, are nevertheless able to cooperate, form a majority, and govern.

---

[9] In his (1896) *Symbolic Logic*, Carroll states, "I maintain that any writer of a book is fully authorized in attaching any meaning he likes to a word or phrase he intends to use. If I find an author saying, at the beginning of a book, 'Let it be understood that by the work "*black*" I shall always mean "*white*", and by the word "*white*" I shall always mean "*black*", I meekly accept his ruling, however injudicious I think it." The difference from the Humpty Dumpty case is that this hypothetical author tells his audience beforehand. He is at least being somewhat communicatively cooperative, though he could have been more so.

[10] The latter option is an interesting possibility. If he meant to say something that he believed Alice couldn't have understood, then Humpty Dumpty succeeded in realizing that intention, and did so with being communicatively cooperative. But in that case he didn't have a *communicative* intention either.

Our second task in support of the Communicative Cooperation Principle is to justify it as a rational behavioral strategy according to our novel theory of moral psychology as outlined in the previous chapter. According to Rational Virtue Theory, any repeatable, stable behavioral strategy that maximizes one's likelihood of achieving his or her ends is rational. CommCoPrin assumes that he one has some communicative intension(s). Whether or not having such intention(s) is rational is a separate question from the validity of CommCoPrin, so we can set it aside.[11] The question is what are the best means for realizing these communicative ends. The best way to achieve that is to cooperate in a very general way with DPs so that they can come to recognize your communicative intentions: they come to understand your intended meaning. CommCoPrin is not specific about how to cooperate so that you're more likely to realize your communicative intentions. For that we need specific maxims.

Luckily, Grice has already provided several.[12] His only mistake was that he regarded the maxims as falling out of ConvCoPrinc (which is roughly equivalent to his Cooperative Principle), which leads us to our third task. As we already saw in the case of the silent response, it is still possible to conversationally implicate something by means of the maxims without being conversationally cooperative.[13] Instead, I argue they derive from Communicative Cooperation: you observe the Gricean maxims in order to increase the likelihood of realizing your

---

[11] That said, it does seem reasonable to assume that having a communicative intention is rational, as it can, and often will, be a good instrumental end for achieving something else. However, it does not follow that all communicative intentions are rational – there may be some things it is best not to express – but many are.

[12] Or at least some of them: see chapter 5, section 2 for discussion of the role of additional maxims in conversation and the conveyance of meaning through implicature. We can also add linguistic maxims about using a grammatical structure and vocabulary that one's DP is likely to comprehend.

[13] To conversationally implicate *by means of the maxims* is to implicate in such a way that one's audience can work out what was implicated by assuming the maxims are being observed.

communicative intentions. And any violation of the maxims should not come at the expense of being less communicatively cooperative. We need not go through to full list of Gricean maxims to see how observing each increases the likelihood of realizing your communicative intentions; a few examples will suffice to make the point. For example, the second Maxim of Manner – avoid ambiguity – is intuitively rational. The more ambiguity in your utterance, i.e., the more possible interpretations (particularly those that are likely to occur to your DPs), the less likely your DP is to pick the intended one as what you meant. And the rationale for lessening ambiguity in no way requires that you be conversationally cooperating with your DP. You might have the communicative intention of insulting your DP, a conversational purpose or direction to which they would in no way accept and share with you. Nevertheless, if you intend for them to understand that they have been insulted, you don't want to phrase the insult in a longwinded and obscure manner that would leave your DP befuddled as to your meaning. Similar intuitive derivations for the rest of the Gricean maxims can be given.

3 Conclusion

Cooperation is a rational behavioral strategy, one that is typically in one's best interest, at least in light of existing social norms and enough others cooperating. Cooperation in various forms also has a clear evolutionary developmental history. Humans at some point learned to cooperate in a way that our closest primate ancestors did not. Children display a wide variety of cooperative behavior from a very early age. Therefore, cooperation seems a likely candidate for inclusion in our moral psychology manual.

Cooperation in language is something distinct. We cooperate not because we necessarily share common purposes or directions with others, though we may. Rather, the justification for cooperation is out of expediency of realizing our own ends. If I want to communicate something, I must do it in such as way that you will understand. And so I cooperate with your expectations of how language should be used.

Chapter 6: Implications and Entailments

In the last chapter I argued for the moral nature of cooperation in general and in communication, as well as Grice's conversational maxims. Grice expected this connection between moral psychology and language. However, some criticisms and corrections of Grice's theory of implicature become necessary now that the connection has been spelled out in greater detail. Certain elements of his theory will require adjustment in light of their new moral status. In this chapter, I will explore some of these theoretical implications. First, I will argue that moral maxims can play the same role in the working out of non-conventional implicatures as Grice thought the conversational maxims do. This result will lead to a new taxonomy of the types of implicature different than (or from) what Grice originally thought. Second, I will present a new analysis of one of Grice's most well-known examples of implicature, the letter of recommendation. I will argue that moral maxims play a necessary (and previously overlooked) role in explaining how the implicature is worked out in this case. Third, I take on several of Grice's detractors, arguing

that their criticism has mistakenly identified failures of speakers as failures of Grice's theory. This argument will prompt a new take on indeterminate implicatures and an explanation of the necessary conditions for successful implicature.

1 Conversational and Moral Implicatures

As I argued in chapter 3, Grice has his four conversational maxims, which play a role in how conversational implicatures are calculated. These weren't the only types of maxims that Grice mentions, however. He admits that there are "all sorts of other maxims (aesthetic, social, or moral in character)" (1975a, pg. 47), though he says little else about them. Given the connection between conversational maxims and conversational implicature, it seems reasonable to interpret Grice as allowing for the possibility of other sorts of implicature, including aesthetic, social, and moral varieties.

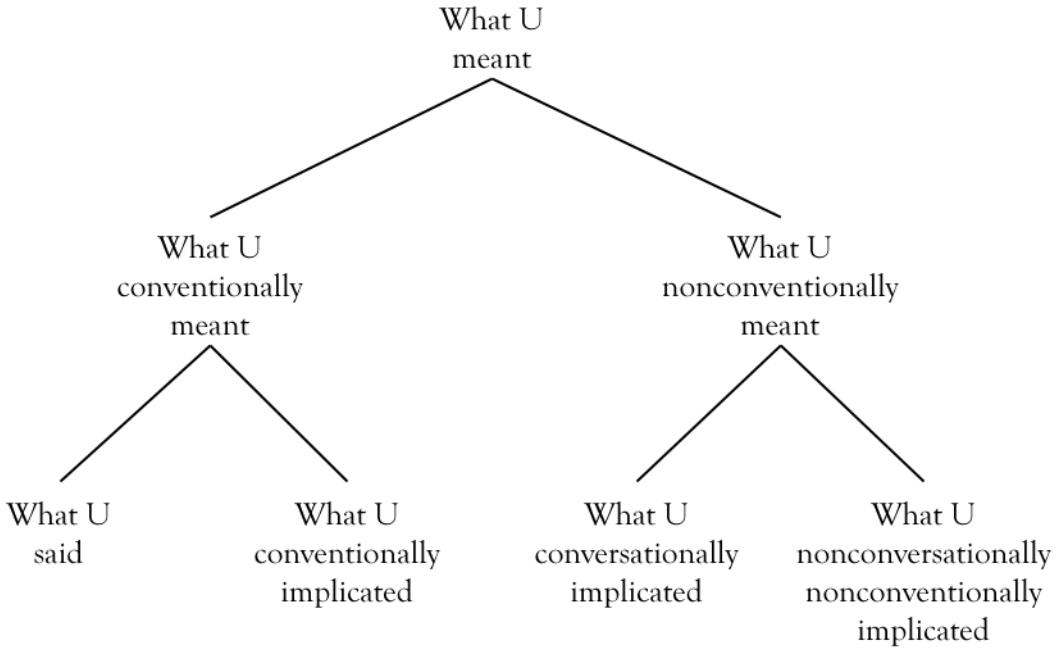
Nevertheless, conversational implicatures (and the related conversational maxims) are supposed to be unique among the various types of implicature because they bear a unique relationship to the Cooperative Principle. In this section, I will present my main criticism of Grice's theory of implicature. I will argue three points: (1) that moral maxims cannot be separated from conversational maxims, at least not in the way Grice supposes, since they relate to implicatures in the same way as the conversational maxims; (2) that a new taxonomy of non-conventional implicature is needed, which I will present; and (3) that the addition of moral maxims (and other non-conventional, non-conversational maxims) forces us to recast most (if not all) cases of non-conventional implicature Grice sees as involving the flouting a maxim as cases involving two or more maxims clashing.

1.1 Background

We need to review briefly Grice's taxonomy of implicature and his discussion of how conversational implicature is classified.

1.1.1 Taxonomy

Neale (1992) presented the now standard interpretation of Grice's theory with this breakdown of a speaker *U* meant with any given utterance.[1]



What *U* meant divides into "What *U* conventionally meant" and "What *U* nonconventionally meant," each of which will divide in two again. What *U* conventionally meant is composed of "What *U* said" and "What *U* conventionally implicated." What *U* nonconventionally meant is

[1] Neale notes that he has omitted the distinction between generalized and particularized conversational implicature, "because it is theoretically inert (for Grice.)" (1992, pg. 524) I will adopt this same approach. In what follows, I'll be dealing with particularized conversational implicature.

sub-divided into "What *U* conversationally implicated" and "What *U* nonconversationally, nonconventionally implicated." The end result is a balanced tree of three levels, symmetrical with four nodes at the bottom.

So according to Grice, there are two main types of implicature. First, conventional implicature is signaled by the conventional meaning of the words uttered. For instance, utterances of "Are you still sleeping with your friend's wife?" and "She's a politician, but she's honest" carry conventional implicatures (respectively that the audience has previously been sleeping with the friend's wife and that the usual contrast between politicians and honesty does not hold in this case).[2] The force of the implicature is provided by the conventional meanings of 'still' and 'but.'

Second is non-conventional implicature, which does not rely on of conventional devices that mark the presence of an implicure. This class is sometimes mistakenly identified with the class of conversational implicatures. The presence of a conversational implicatures is explained by reference to the Cooperative Principle and the conversational maxims, which we will discuss in the next section. As Grice notes (1975a, pg. 70), a conversational implicature *can* be understood in terms of an argument (or at least a gloss of one) that explains, by reference to the maxims, how the speaker was able to conversationally implicate (or might have implicated) that *q* by saying (or making as if to say) that *p*. The conversational implicature must be derivable. Otherwise the implicature is conventional in nature. As Grice puts it,

> The presence of a conversational implicature must be capable of being worked out; for even if it can in fact be intuitively grasped, unless the intuition is replaceable by

[2] As I point out, the presence and the shape of the implicature is signaled conventionally, but the precise content is conversational.

> an argument, the implicature (if present at all) will not count as a conversational implicature; it will be a conventional implicature. (1975a, pg. 50)

Following Neale, I call this the **Justification Requirement** (1990, pg 78).[3]

As Neale's taxonomy recognizes, however, there is a second sub-class of non-conventional implicature besides conversational implicature, a point often missed in the literature. In a brief aside Grice states,

> There are, of course, all sorts of other maxims (aesthetic, social, or moral in character), such as "Be polite," that are also normally observed by participants in talk exchanges, and these too may also generate nonconventional implicatures. The conversational maxims, however, and the conversational implicatures connected with them, are specially connected (I hope) with the particular purposes that talk exchanges (and so, talk exchange) is adapted to serve and is primarily employed to serve. (1975a, p. 47).

Typically, commentators just acknowledge this passage and move on.[4] I want to focus on this other class of non-conventional implicature and see what effect on Grice's theory a fuller account of them would have.

---

[3] Davis (1998) calls it the Calculability Assumption. The title alone betrays his misunderstanding of Grice's view. As Sperber and Wilson (1986) note, in Grice's original formulation, what is being implicated is not logically derived or mathematically calculated, since it is introduced without explanation instead of from a prior premise.

[4] See Neale (1992), Saul (2002), and Davis (1998).

Constituting the class of non-conventional implicature, we already have conversational implicature. I noted above there is good reason to interpret Grice as also admitting moral, social, and aesthetic implicatures, given that he explicitly mentions maxims of this type.[5] Interestingly, there is no argument given to limit the number at four. Perhaps there are legal, political, medical, or familial maxims, and so implicatures of these kinds as well.[6] Regardless of how long the list of types of maxims (and so associated types of implicature), Grice divides them into two groups, conversational and everything else. The distinction Grice has in mind between maxims of the conversational variety and all other maxims turns on their relationship with the Cooperative Principle:

> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. (1975, pg. 67)

Grice's supposition appears to be that his conversational maxims fairly exhaust the nature of cooperation and that one can still be cooperative regardless of whether or not one observes the other maxims. So, while being polite in a talk exchange may be a morally good thing for you to do, it does not make you more or less cooperative towards the "accepted purpose of the talk exchange." My conjecture is that this distinction between conversational maxims and all other

---

[5] I will largely set social and aesthetic implicatures and maxims aside, as they are not central to my project, though some mention of them will be made below in section 1.3. It is also not clear that these must be necessarily distinct categories.

[6] If there were such non-conventional implicature, they would likely be derivative. My point, however, is that no argument is given as to why these four types.

maxims is spurious. Conversational maxims are also moral, and the non-conversational maxims also pertain to communicative cooperation.[7]

Of non-conventional, non-conversational implicature, Sadock (1978, p. 282) only states that he does not see why non-conventional, non-conversational maxims do not relate to the Cooperative Principle in the same way as the conversational maxims do, for not being polite in conversation is also not cooperative. He does not develop this idea, unfortunately, but it is worth further consideration.

Leech (1983), following Grice's line of thought, proposes a Principle of Politeness, admonishing us to avoid communicative discord and strive for communicative harmony. Under this principle, Leech posits six categories of politeness maxims: Tact, Generosity, Approbation, Modesty, Agreement, and Sympathy. For example, the two Modesty maxims state, "Minimize the expression of praise of self; maximize the expression of dispraise of self" (Leech, 1983). While Leech appears to be on the right track with the Principle of Politeness by realizing the importance of politeness in communication, his view is not without problems. First, his view preserves Grice's distinction between conversational and all other maxims. Second, his maxims are too strong, requiring for instance criticism of oneself as much as possible within the confines of other maxims. That is too extreme and simply not observed in linguistic behavior. Third, while politeness in communication (broadly construed) may be common in many cultures – as Leech argues in his (2005) – the maxims are too fine-grained and specific to capture the widespread but diverse application of politeness. Some of Leech's maxims are adhered to by some cultures, but not

[7] I will be using CommCoPrin from the previous chapter instead of Grice's original Cooperation Principle.

others. Fourth and finally, Brown and Levinson (1987, pg. 4) argue that Leech has too many maxims (to say nothing of the possibility for still more), and these are based upon mere communicative regularities. They contend, if there were maxims for every regularity in language use, we would be awash in such a plethora of maxims and pragmatics rather useless.

This final criticism by Brown and Levinson makes a significant error. They are objecting as linguists to the notion that there are moral maxims. There is ample reason to suppose that there are moral maxims for our behavior in general.[8] A more reasonable objection on their part would have been to claim that these maxims do not pertain to our use of language, in particular in such a way so as to be able to generate implicatures. Let us take as a given that there are moral maxims. What is needed, and neither Grice nor Leech sufficiently provided, is a principled argument demonstrating the role of these maxims in the production of implicatures. Grice assumed they did.

1.1.2 Classifying Conversational Implicature

According to Grice (1975), cases of conversational implicature fall under one of three types: **Maxim Adherence** (or no obvious maxim violation), **Maxim Clashing**, and **Maxim Flouting**.[9] In cases of Maxim Adherence, the speaker can be taken to adhering to all the conversational maxims. For example, when A mentions that Smith doesn't seem to have a

[8] See the previous chapters for discussion of the source and legitimacy of moral maxims in general and cooperation specifically.

[9] The first term is my own. Grice does not state that these three ways are the only ways that particularized conversational implicature can be generated. That said, we can limit ourselves to these three.

girlfriend at present, B responds, "He has been paying a lot of visits to New York lately." B would be infringing upon the maxim of Relation unless he thinks that Smith is likely to have a girlfriend in the City.

In cases of Maxim Clashing, a speaker is in a situation in which he cannot satisfy all the maxims with a given utterance. For instance, Grice requires that speakers give as much information as is required by the purpose of the talk exchange. But speakers must also not state that for which they lack evidence. So if asked where a mutual friend lives (when planning a trip to France together), the utterance of "Somewhere in the South of France" carries the implicature that the speaker is unable to provide more information.

The final type of conversational implicature is maxim flouting. Cases of this type involve exploitation: the speaker expects the audience to realize that he is intentionally violating a maxim, but nevertheless making a cooperative conversational contribution. Arguably the most well-known example of an implicature of this type, the letter of recommendation, I will set aside for later. In another example from Grice (1975, pg. 55), the speaker is writing a review. Instead of writing "Miss X sang 'Home Sweet Home,'" the author writes, "Miss X produced a series of sounds that corresponded closely with the score of 'Home Sweet Home.'" The author intends for his reader to realize that he opted for this seemingly unnecessarily long-winded replacement for 'sang' not to be intentionally uncooperative, but to communicate a difference between singing and what Miss X did, presumably as a disparagement to her performance.

1.2 Some examples

To dissolve Grice's hypothesized distinction between conversational and moral maxims (and implicatures) I must demonstrate the following: (1) that moral maxims relate to the Cooperative Principle in the same way as conversational maxims, (2) that they relate to implicatures in much the same way, and (3) that moral implicatures satisfy the Justification Requirement. To that end, consider the following three conversations.

1.2.1 Unhelpful Insults

The first conversation is between speakers A and B, where A is in an immobile car.

Conversation 1:

A: I am out of gas.
B: There is a gas station around the corner, moron*!*

Let's assume for the sake of argument that one of the moral maxims is "Be polite." Then *B* has violated that maxim. Conversation (and communication generally) is not a special type of human behavior in which moral maxims and norms have no purchase, as Brown and Levinson appear to suppose. In *Conversation 1*, *B*'s addition of "moron" was clearly not polite. But more importantly, it was not fully cooperative.[10] Insults were never part of the accepted purpose or direction of this talk exchange. *B* is refusing to be fully cooperative, likely implicating something about his frustration either at having been asked a question he regards as asinine or taking up time he would prefer to spend otherwise. Moral maxims pertain just as much to the Communicative Cooperation Principle as the conversational maxims do. One might think that *B*'s response is not cooperative because of the extra and irrelevant addition of 'moron," which is prohibited by the second maxim of Quantity. A simple change to *B*'s utterance shows that this view is incorrect. Replace 'moron'

[10] *B* cannot be said to be completely uncooperative however, since he did answer the question.

with *A*'s name, still used as a vocative. In this case, the addition of *A*'s name doesn't mean that *B* is uncooperative because he added irrelevant information. So too then if *B* says 'moron.' The lack of cooperation flows from another source.

1.2.2 Polonius's Brevity and Wit

Our second conversation comes from Shakespeare's "Hamlet." Polonius has come before King Claudius and Queen Gertrude to reveal delicate information about the sad state of noble Hamlet. The conversation (Act 2, scene 2) begins thus:

Conversation 2:

POLONIUS

My liege, and madam, to expostulate
What majesty should be, what duty is,
Why day is day, night night, and time is time,
Were nothing but to waste night, day and time.
Therefore, since brevity is the soul of wit,
And tediousness the limbs and outward flourishes,
I will be brief: your noble son is mad:
Mad call I it; for, to define true madness,
What is't but to be nothing else but mad?
But let that go.

QUEEN GERTRUDE

More matter, with less art.

POLONIUS

Madam, I swear I use no art at all.
That he is mad, 'tis true: 'tis true 'tis pity;
And pity 'tis 'tis true: a foolish figure;
But farewell it, for I will use no art.
Mad let's grant him, then: and now remains
That we find out the cause of this effect,
Or rather say, the cause of this defect,
For this effect defective comes by cause:
Thus it remains, and the remainder thus. Perpend.
I have a daughter–have while she is mine–
Who, in her duty and obedience, mark,
Hath given me this: now gather, and surmise.

Polonius goes on to read a letter from Hamlet to Ophelia (with comments), describing Hamlet's affection, and then concludes that Ophelia's spurning Hamlet's advances (at Polonius's behest) has driven Hamlet to madness. We don't need to quote further the passage in its considerable entirety to realize a key point, namely that Polonius takes a long time to make his point, something the Queen remarks upon with her "More matter, less art." Hamlet's madness and its cause is a direction for the talk exchange that Polonius introduced himself. So we, and his audience, must wonder why he is being so long-winded, displaying such (seemingly) unnecessary prolixity, and if there is some communicative intention behind his plethora of verbiage.[11]

Suppose there is a moral maxim at play, something of the form "Be mindful of people's feelings" or "Don't discuss impolite or sensitive topics." (Or "Put delicate news delicately.") The precise formulation of the moral maxim is not important at the moment, so long as it is clear that there is some maxim at play prompting speakers to lessen the negative emotional impact to the audience.[12] A common way to achieve this end is to preface the remark or make excuses for the

---

[11] A possible explanation is that Polonius is simply a fool, a clown Shakespeare introduces simply to mock, but not to be taken seriously. As such, there would be no communicative intention underlying Polonius's loquaciousness. Not to get lost in Shakespeare criticism, even if he is properly regarded as a comic figure, we can imagine the same conversation being uttered in earnest by a serious twin of Polonius. Then his pleonasm becomes meaningful. That is the conversation I am focusing on above, regardless of the correct interpretation of Shakespeare.

[12] The justification for such a maxim primarily comes from concern for other people (either because doing so is an end one has adopted or because doing so is a game-theoretically justified strategy that is typically in one's self-interest to follow). But there can often be an immediate self-interested rational for adhering to this maxim, namely to avoid a 'shoot-the-messenger' situation or to exonerate oneself from any blame for the news. The latter appears to be Polonius's main motivation.

news.[13] A messenger tasked with telling the king the news of the army's defeat might preface his remarks with mention of all the mitigating circumstances and that the loss occurred despite the gallant efforts of so-and-so and the noble sacrifice of such-and-such. A similar phenomenon is found in good news/bad news cases.

Doctor: "I've got good news and bad news."
Patient: "Give me the good news first."
Doctor: "We're naming a deadly disease after you…"

The doctor is trying to lessen the blow. It is likely that something similar is at work with innuendo or other roundabout expression for socially awkward or taboo topics.[14]

If this moral maxim is at work, Polonius's utterance makes a great deal more sense. Without question, he has violated the third maxim of Manner ("Be brief"); he has engaged in extreme prolixity. But is there any implicature is present? If Polonius has simply violated the maxim of Manner and also (partially) opted out of the Communicative Cooperation Principle, there would be no implicature. Still, why would he introduce a direction for the talk exchange (what's wrong with Hamlet) and then refuse to cooperate? There are only two plausible solutions: either he is an incompetent speaker or he meant something by speaking as he did. Just as Grice did (and audiences regularly do), we can assume that the speaker is not incompetent. Some implicature must be present, then. There are two maxims in tension: the conversation maxim of Manner and the moral maxim of politeness. Polonius cannot fulfill both simultaneously. His infringement of the third maxim of Manner can then be explained by the supposition that he

---

[13] It is debatable how successful such behavior is, much like the notion of pulling off a bandage slowly will be less painful. Regardless, speakers regularly employ this strategy.

[14] See Monty Python's "Candid Photography" sketch (otherwise known as "Nudge, Nudge"), in which Eric Idle uses a series of obscure innuendo followed by "Nudge, nudge, wink, wink, say no more, say no more" to get at the *bawdy* topic of sex.

recognizes less prolixity would infringe on the moral maxim. Hence, Polonius appears to be implicating something about his recognition of the delicate nature of his news.[15] Note that we were able to construct an argument with reference to the Cooperative Principle to explain the presence of the moral implicature, thus fulfilling the Justification Requirement.

1.2.3 The Weather

The third conversation comes directly from Grice (1975, pg. 54). Consider the following, which occurs at a "genteel tea party."

Conversation 3:

A: Mrs. X is an old bag.
B: The weather has been quite delightful this summer, hasn't it?

This is one of Grice's examples of Maxim Flouting, specifically the maxim of Relation. *B* is refusing to make a remark relevant to the direction of the talk exchange set by *A*. Grice reasons that since *B* has so blatantly violated a conversational maxim, *B* must be implicating that *A*'s topic is not to be discussed or that *A* has committed a faux pas.[16]

---

[15] It is also possible that his implicature has illocutionary force, since his intention is not just to convey information about his awareness that this is likely painful news (at least to Gertrude, and maybe to Claudius), but also to achieve a particular result, namely to alter the emotion state of his audience from the worse state it would be if he spoke directly.

[16] Though Grice does not mention it, I think this is a case of generalized conversational implicature. There is something of a convention built up around mentioning the weather as an obvious change of topic. It would not have worked if *B* had said, "Barack Obama is the President, right?" or "Red is a color, correct?" Or at the very least, they are less likely to convey the implicature than mentioning the weather. Such remarks would more likely simply seem odd, so that any meaning about the inappropriateness of *A*'s comment would likely be lost. Some non sequiturs are just that, but a few others – like references to the weather – typically convey in this English-speaking culture an implicature. Though based on convention, the implicature still relies upon the Cooperative Principle and violating the maxim of Relation, since not all references to the weather conventionally carry implicata of this sort. The point is that a convention can develop about how to violate a maxim in a meaningful way in order to signal the significance of that

I contend that rather than being a case of Maxim Flouting, there are clashing maxims at work. The problem with thinking that *B* is flouting a maxim (instead of just outright violating it – in which case there would not be any implicature) is that *B*'s conversational contribution would not be cooperative in that case. *A* has set the direction of the talk exchange – to rag on Mrs. X. *B* is refusing to cooperate. If *B* is not only violating a maxim, but opting out from cooperating with the direction of the talk exchange all together, there can be no conversational implicature.

Nevertheless, *Conversation 3* is an instance of a conversational implicature. But to demonstrate that fact, we require an explanation of how *B* is being cooperative and what argument can be constructed whereby *B*'s meaning becomes clear (in order to fulfill the Justification Requirement). To see the problem with the flouting view, consider how the audience comes to understand *B*'s meaning. How does one know, for instance, that *B*'s meaning is about the inappropriateness of *A*'s comment instead of implicating something about how *B* thinks the issue of the weather is much more important or urgent than discussion of Mrs. X? *B* might have said, "The house is on fire!" This exclamation is also not cooperative with the direction *A* set. But in this version, there is no conversational implicature about *A* having committed a faux pas.

How then is the audience to work out that when mentioning the weather, there is a conversational implicature (and what is implicated), but that there is not one when warning of fire? Some additional information must either be provided or assumed in order to work out *B*'s

---

violation. This conventionality, however, does not mean that the implicature is conventional instead of conversational – contrary to Davis (1998). Finally, what is especially interesting about this cases of generalized conversational implicature is that it is the subject matter – the weather – that has become generalized, instead of a particular word or phrase. The point would have been equally clear had *B* said instead "We've had just the perfect amount of rain lately, don't you think?" or "Cloudy out, huh?"

intention, and so his meaning. A possible source could be the context of the conversation, which Grice states is a "genteel tea party." The idea, then, is that gossip and speaking ill of someone is not acceptable in this a context, though it might be in others.[17] Given this context, we can interpret *B* as reinforcing the constraints or rules of the context ("Be genteel" or more specifically "Don't gossip or speak ill of someone"). The problem with this view is that the same context applies to *A*'s comment as well as *B*'s. We must interpret *A*'s comment in light of this context, meaning that *A* would have to be implicating something about his disdain for the genteel tea party or its rules; he was being intentionally crass. Just as much as referring to the context does give us a plausible reconstruction of *B*'s meaning, so too it could be the case that *A* was intentionally flouting the conversational constraints of the context. But it is by no means necessary. While referring to the context can be helpful to the audience, there is however a more reliable alternative.

A better alternative is to make some assumptions about *B*'s moral psychology. Suppose the moral maxims recited by a myriad of mothers, "If you can't say something nice, don't say anything at all" is valid. Here we have a moral maxim that only pertains to communication, further eroding the notion that moral and conversational maxims are distinct. *B* now faces a clash of maxims: he cannot be relevant and nice at the same time. If *B* is flouting the maxim of Relation, it is to signal his recognition that the conversational maxim clashes with a moral maxim. Recognition of the clashing maxims – and which maxim *B* chose to observe – tells the audience what *B* is implicating.

[17] Neale has recently presented a number of problems with appeal to context in determining a speaker's meaning (2005 & 2007).

So we have fulfilled the Justification Requirement for a conversational implicature. If we look to moral psychology, the answer becomes clear.[18]

1.3 Revisions

These findings about moral principles and their role in working out non-conventional implicatures require that some revisions to Grice's theory be made. Contrary to Grice's stated view, there is no principled way to divide the Communicative Cooperation Principle from the moral maxims that pertain to language use. The main revision that is needed then is a new taxonomy of implicature. This new classification will also require a modification of the Justification Requirement.

Recall that in Neale's taxonomy, there are two types of non-conventional implicature: conversational and everything else (non-conversational, nonconventional implicature). I have no argument against Neale's interpretation of Grice. Rather, my argument is with Grice himself, particularly regarding the *What U non-conventionally meant branch*. Though it will break the pleasant symmetry of Neale's diagram, my revision will re-define conversational implicature so that it is the only node beneath "What U non-conventionally meant." The category of "What U non-conversationally, non-conventionally implicated" will be subsumed under conversational implicature, because we found no basis for the distinction.

We already have the Communicative Cooperation Principle and the conversational maxims. Grice also spoke of moral, social and aesthetic maxims (though without any mention of

---

[18] One final point on why we can't appeal to context to provide the implicature: this is only a genteel tea party because its participants (or enough of them) intended it to be. The context, just like meaning, is determined by the intentions of those involved. So appeal to context in this case is still looking at the speaker's intentions to provide the meaning.

corresponding principles). Since Rational Virtue Theory posits moral principles and the Communicative Cooperation Principle was seen to be such a principle, I suggest we posit moral, social, and aesthetic principles under which the corresponding maxims reside. We already have the example of the Politeness Principle with the corresponding maxim of "Be as polite as required." That may be the only maxim under the principle, though other moral principles may have multiple maxims under them, just as the Communicative Cooperation Principle does. Further, all these principles should be covered under the heading of the **Conversational Principles**.

A **Conversational Principle** is any normative principle pertaining to communicative behavior that has been (or for now, is likely to be) game-theoretically justified as a rational, *ceteris paribus* behavioral strategy that, when adhered to, increases the likelihood of realizing one's higher-order end(s). The Communicative Cooperation Principle is our prime example of a Conversational Principle. As we saw in the previous chapter, behaving as the Communicative Cooperation Principle prompts us is a good idea. Doing so typically increases the likelihood of achieving some of our other end(s), such as being understood by others, developing relationships with others, or receiving information. As I conceive of it, what principles are in the set of Conversational Principles will largely be stable across cultures. Cooperation and politeness are generally in one's self-interest across cultures. However, the precise maxims or application of those maxims can differ across cultures. What counts as being polite in the United States today is different from being polite in Elizabethan England or during the height of the Tokugawa

shogunate in feudal Japan.[19] We have removed the distinction between conversational implicature and non-conversational, non-conventional implicature. Besides the Communicative Cooperation Principle, any moral, social, or aesthetic principle that pertains to communicative behavior falls under the heading of Conversational Principles. Therefore, any non-conventional implicature is a conversational implicature. There is nothing left over.

With the addition of Conversational Principles, we must also re-formulate the Justification Requirement. Grice states that a conversational implicature must be capable of being worked out by the audience, and they can do so by presuming that speaker is observing the conversational maxims, or at least the Communicative Cooperation Principle.[20] Now, the audience must *be able* to work out what was conversationally implicated by reference to all Conversational Principles.[21] In fact, as we have just seen in the recent examples, access to additional normative principles, such

---

[19] One might suggest that the Politness Principle be something like the following: "Be as polite as is required," since different conversations require varying levels of politeness. So politeness is a quantitative matter along a sliding scale. This formulation then follows Grice's example of including "as is required" in the two Quantity sub-maxims. But none of the other sub-maxims follow the same example. Grice states we should "be orderly" and "avoid ambiguity," instead of "be as orderly as required" and "be as unambiguous as required." And with good reason. Puns are intentionally ambiguous and must be so by necessity. They also flaunt the maxim prohibiting ambiguity in order to make the ambiguity and the joke obvious. If the sub-maxim read "be as unambiguous as required," one wouldn't be flaunting the maxim when punning.

Likewise with politeness, including "as is required" seems to be the wrong move. What even counts as polite can vary from one culture to another. So you must first understand qualitatively what is polite in your current conversational context, and only then determine quantitatively how much politeness is required.

[20] The naming of maxims and principles can become slightly confusing at this point. The conversational maxims, which stem from the Communicative Cooperation Principle alone, have long had their cognomen, and I will not attempt to rename, despite the confusion, though 'cooperation maxims' would now be a better title. They are not the only maxims that fall out from the Conversational Principles, since there are other moral, social, and aesthetic maxims.

[21] I do not mean that the audience would have to make use of every Conversational Principle, but that they could use any of them.

as the Politeness Principle, will allow the audience to work out what was meant more precisely than what they could do with only the Communicative Cooperation Principle. That said, however, it seems likely that among the Conversational Principles, the Cooperative Principle will take precedence just as the maxim of Quality does among the Gricean maxims. A speaker's first obligation is, *ceteris paribus*, to be cooperative. If he is not doing that, any failure to be polite (for instance) is secondary.

2 The Letter of Recommendation

The most well known example of Maxim Flouting is the one subsequently named for Grice, the Gricean letter. In this vignette, Prof. A is writing a letter of recommendation for Mr. X, one of his students, who is applying for a philosophy job. Prof. A writes, "Dear Sir, Mr. X's command of English is excellent, and his attendance at tutorials has been regular. Yours, etc." (Grice 1975, pg. 52). Grice summarizes his analysis of this letter as follows:

> A cannot be opting out, since if he wished to be uncooperative, why write at all? He cannot be unable, through ignorance, to say more, since the man is his pupil; moreover, he knows that more information than this is wanted. He must, therefore, be wishing to impart information that he is reluctant to write down. This supposition is tenable only if he thinks Mr. X is no good at philosophy. This, then, is what he is implicating. (1975, pg. 52)

Grice's analysis rests on three conjectures:

1) The fact that Prof. A is writing the letter at all demonstrates that he is being cooperative;

2) The only possible way for the Prof. A to be providing relevant information is that he must mean something he is unwilling to say (or write down, as it were);

3) A must mean that Mr. X is not good at philosophy.

I find all three of these suppositions misguided or problematic. Let's grant that a conversational implicature is present. The main problem is that Prof. A's audience could not work it out in the manner that Grice describes, and (consequently) it does not appear to be a very cooperative or well-chosen implicature.[22]

2.1 Cooperative Assumption

I will discuss the problems with each of Grice's three conjectures in turn. Conjecture (1) states that the fact that Prof. A wrote the letter indicates that he is being cooperative. This seems an odd notion. There are really two claims that are worth dividing:

(1a) Prof. A is being cooperative.

(1b) The audience *knows/can realize* (1a) from the fact that A wrote the letter.

First, consider (1a). How is he being cooperative with this letter? He certainly is not cooperating with the student, Mr. X, who presumably had requested a recommendation, of which this is anything but.[23] The conversation, however, is really between Prof. A and the faculty member(s) at

[22] A well-known philosopher (who shall remain nameless) recently remarked to me, "We all write Gricean letters." It seems likely that since Grice wrote this, the formula for a Gricean letter has become well known and generally recognized throughout our discipline – perhaps more so than in other disciplines – such that the implicature has become more conventional for philosophers than Grice's analysis supposes.

[23] It has always troubled me that A would have agreed to write for X in the first place if he was not willing to recommend him. There seems to be some moral failing by A for agreeing recommend X and then not doing so. (Perhaps I'll feel differently later in my career.) It seems likely that much of the oddity and confusion about this example stems from this fact. The reader is expecting not

the school(s) to which Mr. X has applied (whom we can call R). What then is the accepted purpose or direction of this talk exchange (taking 'talk' loosely)? If the purpose is to recommend Mr. X, then there appears no sense in which A is cooperating. But if we take the purpose to be the discussion of Mr. X's fitness to teach philosophy, then some inferences can be made as to A's meaning. Interesting to note is that is (at least) one other option for the purpose of this talk exchange and A's letter. Instead of being about X's ability as a philosophy instructor, it could be about X's quality as a philosopher. The conclusion R reaches about A's meaning will depend on the purpose. But which is it, and how is R supposed to know which?

Next comes (1b). What about the fact that A wrote the letter entitles the audience to conclude that A is cooperating with the purpose of the talk exchange (whatever that is)? Despite Grice's conjecture, the fact that A wrote the letter at all does not necessarily warrant the conclusion by R that A is being cooperative. It might have been the case that Prof. A, as an older member of the faculty, has become a bit confused in his later years and wrote a letter that just did not make much sense. Or perhaps, unknown to X, Prof. A holds a grudge against X, who really is a good philosopher, but A wants to take his revenge with this letter. Both of these are uncooperative possibilities that include A not only writing a letter, but this same letter. Hence, R cannot conclude that A must be cooperating simply because A wrote.

Conjecture (1b) cannot be countenanced. But it does not follow *ipso facto* that R should conclude A is being uncooperative. Rather, that A is being cooperative is a necessary working assumption, a prerequisite for interpreting A's meaning (on the assumption there is any). Let's call

---

just a letter, but a letter of recommendation. And there is no way to construe A's meaning as any form of a recommendation.

this the **Cooperative Assumption**: *ceteris paribus* the audience should initially assume as a default position that the speaker is being cooperative, absent prior evidence to the contrary.[24] This assumption is only an initial starting point for the audience. If after consideration, the audience cannot decipher the speaker's meaning, then at that point the assumption must come under fresh review. So, as Grice suggests, we can take Prof. A to be cooperating – but not for the reason Grice proposes – and accept (1a) for now. The matter will come up again, however.

2.2 Clashing, not Flouting

Grice's second conjecture (2) is about the reason for the peculiar letter. Let's divide it into:

(2a) There must be something that A is unwilling to say (write down).

(2b) Whatever the contents of (2a) are, that is what A is trying to convey with this letter.

R has already granted (for now) that A is being cooperative, and therefore knows that more information than A provided is required, given the purpose of this talk exchange. The only means by which R can arrive at A's meaning is by determining *why* Prof. A has violated the first maxim of Quantity. For A to be cooperative, it must be that his violation of the maxim is purposeful and flagrant, i.e., obvious enough to be noticed. A wants his audience to wonder why he is violating the maxim and thereby discern his meaning. But the reason behind the flagrant violation is not obvious from the fact that the violation was flagrant. More information about A's intentions is needed. The purpose of the talk exchange is not informative enough.

[24] The Cooperative Assumption can also be part of our theory of moral psychology. See Davis (1998) and his similar Cooperative Presumption.

What reason then could prompt a rational and cooperative speaker to intentionally and obviously violate a conversational maxim? Whatever the reason is, it cannot be a private reason discernable by no one else, if he is to be perceived as cooperative. The audience must be able to conceive of it as a possible reason motivating the speaker's action. No doubt, (2a) is a plausible reason. But the question is what justifies R in concluding that (2a) is true. Where does this plausibility come from? The conversational maxims are components of any rational and cooperative speaker's moral psychology, which we are assuming applies to A as well. If there is another maxim that the audience can expect to a rational speaker to be observing and would also explain A's utterance, then the supposition that A is observing some other maxim could provide an explanation. So R can appeal to A's moral psychology to provide such a maxim. Let's suppose that "If you cannot say anything nice, do not say anything at all" (or something reasonably similar) is a legitimate moral maxim, something rational agents should observe. It is this moral maxim that grants plausibility to (2a).

This example is no longer one of flouting a maxim, but of two maxims clashing. Prof. A cannot observe both this moral maxim and the first maxim of Quantity.[25] The presence of a second maxim clashing with the first maxim of Quantity provides an explanation as to why A flagrantly violated the conversational maxim, an explanation absent if this were just a case of Maxim Flouting. The same problem applies to most supposed cases of Maxim Flouting: the reason for the violation of a maxim is not provided solely by the fact that the maxim was noticeably

[25] It is possible (if not probable) that this is a social norm, rather than a universal maxim for all rational agents. The maxim included in the manual is something like "*Ceteris paribus*, observe social norms." So there is still a second maxim at work, though the clash will not necessarily occur in other cultures that do not have this social norm.

violated. Typically, a rational speaker should and will only violate a conversational maxim if required to do so by some other rationally justified maxim in his moral psychology.[26] We saw the same point with *Conversation 3* above. Therefore, it seems likely that almost all cases of conversational implicature by maxim flouting are mis-identified, only telling part of the story. A wider consideration for other maxims at work (moral, social, aesthetic, or others) will explain the reason for the violation of the conversational maxim and indicate what the speaker is implicating.

(2a) is acceptable, though Grice does not explain why. He takes that it simply must be the case that A is unwilling to write something down. As we have seen, it is likely that A is not willing to write something down, but it is only probable given an additional element of A's moral psychology that Grice neglects to mention. Having properly reached (2a), (2b) is now acceptable, since we have seen how regard for the clashing maxim points to what the speaker means.

2.3 What's Implicated

Unlike (1) and (2), (3) does not require any further division. My one complaint against (3) is that again, it does not *necessarily* follow that A means Mr. X is not good at philosophy, contrary to Grice's conjecture. That is a possible meaning for A's utterance, but not the only one. It might well be the case that Mr. X is very good at philosophical thought and writing. He may have published several papers with original theses and arguments. But X might also be the worst instructor of philosophy ever. Assuming that Prof. A knows all this, it would not follow that A

---

[26] In cases of a conversational maxim clashing with another moral maxim, the speaker is not required always to violate the conversational one. He has a choice, and which one he observes and which he violates will be indicative of his moral psychology. I will discuss this point more in the next chapter. One possible remaining type of flouting is obviously violating a maxim to implicate that one does not accept or thinks little of that maxim.

meant to convey to R that X is not a good philosopher, since quite to the contrary he is. In this case, A might mean that Mr. X is not a good instructor and should not be hired.[27] So which does A mean, that X is not a good philosopher or not a good instructor?[28] Referring back to (1) and (2) – that A is being cooperative and not willing to write something – does not help. Without further information, without some additional common knowledge between A and R, R cannot decide definitively between these two possible meanings. So Grice has over-reached in his gloss of what A *must* mean. Nevertheless, there is some clear implicature. R can legitimately conclude that A is implicating that Mr. X should not be hired. Whatever else A means, he is clearly not recommending X for the job. We are left then with the problem of indeterminate implicature. It also prompts us to consider cases in which when an implicature doesn't work, that is, when it isn't understood by the audience. I'll take up both of these issues in the next section.

3 Responses to Some of Grice's Critics

Our new analysis of the Gricean letter left us with two related questions, how to account for indeterminate implicatures and how to account for cases in which a speaker failed to make clear what he was implicating (and so also what he meant). To answer these questions, I will turn to some of Grice's critics. My central contention is that the criticisms we will examine all make the same fundamental error: they mistake errors by speakers for errors of the theory. We'll begin with

---

[27] Note the addition of the normative implicature. Grice does not include it, but I contend it is also implicated on his reading as well. In fact, Grice generally does not examine the possibility of normative implicatures. Section 3 will include some discussion on this point.

[28] These are only two possible implicatures. There are a great many more. E.g., A might be both a good philosopher and instructor, but intolerable to be around, and so A is not recommending him for those reasons.

objections to Grice's notion of audiences being able to work out what was implicated (which Neale termed the Justification Requirement).

3.1 Calculability and the Justification Requirement

Sadock (1978) objects to the Justification Requirement, arguing that calculability fails in two different ways: sometimes conversational implicatures can be worked out when none exist and sometimes too many conversational implicatures are able to be worked out. Euphemisms are an example of the first type of failure of the Justification Requirement. For instance, 'go to the bathroom' was originally used to convey a conversational implicature, Sadock conjectures, but over ftime has become a matter of conventional. Nevertheless, the meaning can still be worked out as a conversational maxim by reference to the Communicative Cooperation Principle and conversational maxims, though there is no longer any need to do the working out. He concludes that in this case Grice's theory has incorrectly predicted the presence of a conversational implicature.

I take this objection to miss the point entirely. The Justification Requirement is a necessary condition, not a sufficient one. Just because a speaker's meaning can be worked out in this way does not mean that a conversational maxim exists. Just because we can explain the meaning of "I have to go to the bathroom" by means of a conversational implicature does not we have to or that there is one present. The Justification Requirement is not a tool for predicting conversational implicatures, at least not in this manner. Conventional implicatures take precedence. So we should first test for a conventional implicature. If one is found, the

Justification Requirement should never come into play. We stop there, never testing for a conversational implicature.

Sadock's second argument against the Justification Requirement focuses on indeterminacy. He contends that uttering "It's cold in here" can conversationally implicate anything from a request to close the door, open the door, fetch a blanket, pay the gas bill, or a host of other propositions. "In fact it's difficult to think of a request that the utterance could *not* convey *in the right context*" (1978, pg. 320, my emphasis). The problem with the argument is that, as stated, it too misses the point. Simply because the same utterance can be used to conversationally implicate different propositions in different contexts is not a problem. That fact only speaks to the flexibility of language. What would be a problem is if on a particular occasion of someone uttering that sentence in a *particular* context the audience couldn't work out or understand what was being implicated. Sadock does not consider such a case. Even in such a case, however, the failure is that of the speaker, not the theory. The speaker would have done a poor job of making his meaning clear.

Davis (1998) extends upon Sadock's arguments. In general, however, Davis suffers from an overly (and perhaps uncharitably) systematized analysis of Grice's view of conversational implicature. This systematized view is decidedly antithetical to Grice's view and approach. Hence, there is a worry that the view Davis is arguing against is not Grice's own view, but rather some straw-man abstraction. Let's set this concern aside, however. Davis argues that Gricean theory is chiefly based on four principles, one of which I will respond to, the Justification Requirement. It is Davis's contention that the four principles of Gricean theory cannot account for the phenomenon of implicature.

Davis presents the case of Mary, a famous philosopher, writing a letter of recommendation for her daughter Jane. It is taken as common knowledge that Jane is Mary's daughter. In her letter, Mary writes, "Jane is not a bad philosopher," without offering further information. Davis rightly concludes that we and the audience cannot determine if this is case of litotes or damning with faint praise. In the former, the implication is that Jane is quite good at philosophy, while the conversational implicature for the latter alternative is that Jane is not good at all. Davis argues that because the implicature is indeterminate, the Justification Requirement is not satisfied. The audience couldn't have worked out what was implicated, so nothing was implicated.

Based on this and similar examples, Davis points to the failure of the Gricean as a theory as a whole.[29] Here, I think that Davis (and Sadock as well) has taken the error of speakers and transferred them to Grice's theory. When the conversational implicata are indeterminate, we can only conclude that Grice's theory has failed if what the speaker meant to convey is clear to the audience, but what was non-conventionally implicated cannot be explained by reference to Conversational Principles, then the theory has failed. If, on the contrary, what the speaker implicated (and meant) is not understandable by the audience, then that is not a failure of the theory, but the speaker. Miscommunication is an everyday occurrence. Any theory that predicted otherwise would be obviously and regularly falsified. In Davis and Sadock's examples, the failures are of the latter type. Specifically, it is not clear to the audience what the speaker meant. These are cases of failure by the speaker to clearly convey their meaning, of unsuccessful conversational implicature.

---

[29] This is by no means Davis's only criticism. His remarks are too extensive for a thorough reply here.

3.2 Indeterminacy and Success for Implicature

Returning to the Gricean letter case, we were able to reach some conclusion about what A was implicating, namely that X should not be hired. But there seems little doubt (and Grice's gloss confirms) that Prof. A meant to implicate more than this. He failed to communicate his full meaning to his audience.[30]

Saul (2002) also explores the idea of failure in conversational implicature (though with a much different result than I am proposing). She develops a variant of the Gricean letter in which she writes a letter of recommendation for John, whom she believes is applying for a philosophy position and she knows to be a poor philosopher and a kleptomaniac. In her letter she truthfully writes that John is an excellent typist, intending for a philosophy search committee to recognize her implicature. However, John is actually applying for jobs as a typist. So, when potential employers read her letter, they do not think there is any implicature and that Saul meant precisely what she said. Hence, Saul concludes there was no conversational implicature and there can be a gap between what a speaker means on the one hand and what a speakers says and implicates on the other.

While Saul did well to introduce an element of success when it comes to conversational implicatures, her conclusions widely miss the mark. Her arguments center on what are commonly

---

[30] The Gricean letter is an interesting case in which the speaker's meaning is not well or fully conveyed, but nevertheless achieves the desired and intended result of preventing Mr. X's employment. Even if the reader could not work out that Prof. A was implicating that X should not be hired (only being able to conclude it is a weird letter), the fact that the letter is not a clear and direct endorsement is likely enough to dissuade the reader from hiring X.

(but erroneously) taken to be three necessary conditions for conversational implicature, as set by Grice (1975). A speaker S conversationally implicates Q by uttering P iff:

i. Audience H assumes that S is observing the Cooperative Principle (*Cooperative Assumption*);

ii. H must suppose that S believes Q in order to for H to believe (i) (*Determinacy*); and

iii. S believes, and expects H to believe, that H is able to work out that (ii) is true (*Common Knowledge*).[31]

Conditions (i) through (iii) are generally taken to constitute a definition of (or at least a set of necessary conditions for) conversational implicature.[32] But this is not correct. Conditions(i) – (iii) are not conditions for the existence of a conversational implicature. They (at least when appropriately modified) are better understood as criteria for *successful* conversational implicature, not.[33]

To see why, we need to be careful about distinguishing between two things, which are far too often conflated. There is a difference between what a speaker means (including what was said and what was implicated) and what the audience takes the speaker to mean (and how they

---

[31] Davis (1998, pg. 13) incorrectly names (iii) "Mutual knowledge." See Myerson (1991, pg. 64–5) for the terminological difference in game theory. 'Mutual knowledge' refers to knowledge that is merely shared between all of a game's participants without it being the case that each knows the others has this knowledge, and each knows the other knows etc. The addition of knowledge about the other's knowledge is common knowledge.

[32] See Harnish (1976, pg. 333), Lakoff (1977, pg 99), Brown and Levinson (1978, pg. 63), Bach and Harnish (1979, pg, xvii, 6), Atlas (1979, pg. 272), Posner (1980), Wilson and Sperber (1981, pg. 160), Levinson (1983, pg. 113), Green (1989, pg. 93–97), Fasold (1990, pg. 131–2), Hirschberg (1991, pg. 16–24), Berg (1991), Blakemore (1992, pg. 126, 137), M. Green (1995, pg. 98), Davis (1998, pg. 13, 17), and Saul (2002).

[33] At present, I am not primarily concerned with interpreting Grice's remarks, but with the correct theory of conversational implicature. Perhaps a reading of Grice can be given that accords with the position advanced here, but I set that aside for now.

ascertain that meaning).[34] Not only are they temporally separate, they belong to different agents. The same goes for what is implicated and determining what was implicated. Conditions (i) through (iii), as stated, run roughshod over this distinction. The first two are requirements of the audience, completely out of control of the speaker. The third, however, oddly shifts back to the speaker and his beliefs. Saul is correct to assert that there are normative criteria governing not only what is meant and said, but also what is implicated. And sometimes these normative criteria are not satisfied. However, she was incorrect to assert that not meeting these criteria constitutes failing to conversationally implicate anything.[35] It is a confusion of metaphysics and epistemology to conflate existence and content of an implicature with what is involved in ascertaining the existence and content of a purported implicature.[36]

Let's take it, then, that a speaker means whatever the speaker intends to mean and implicates whatever he intends to implicate. It is a separate question of how well he did in conveying that meaning, including any implicature. We can modify the conditions above into necessary conditions for successfully communicating a conversational implicature. We have already come across the Cooperative Assumption as a working assumption by the audience about the speaker, which does not pertain to whether or not the speaker actually was attempting to convey anything by means of a conversational implicature. I will hold discussion of Determinacy

---

[34] I take this claim to be in line with the intention view about meaning in general. See Neale (1992), (2005) and (forthcoming).

[35] Neale and Schiffer correctly identify the source of Saul's confusion as based on severe misreading of Grice. Saul fails to recognize that both speaker's meaning and what the speaker implicates are solely based on the speaker's intentions. See Neale (2005 pg. 181-2, n. 30) for more detail on the error of Saul's view.

[36] This is a view Neale has advocated for some time. See his (forthcoming). Bach (2006), Schiffer, and Devitt (2008) have been clear on this point for some time as well.

for a moment, only noting that it is fine as stated. But the third condition will not be adequate. It is not sufficient for success that the speaker has certain beliefs and expectations about the audience. Common knowledge might not hold, despite the speaker's beliefs and expectations (whether legitimate or not). The inference to what the speaker implicated might be more complicated or subtle than the audience is capable or working out. Either way, the speaker has not successfully conversationally implicature anything.[37] Therefore, I propose the following necessary conditions for successful conversationally implicating Q with utterance p:

i. Audience H assumes that speaker S is observing the Conversational Principles when uttering p (*Cooperative Assumption*);

ii. H must suppose that S believes Q in order to for H to believe (i) (*Determinacy*); and

iii. H believes that:

    a. S believes that H can work out that (ii) is true (*Justification Requirement*); and

    b. S expects H to believe that S believes (iiia) (*Common knowledge*).[38]

Like Grice's discussion of calculation, I am not requiring the audience actually to have worked out this process, but that they could make these recognitions. These success conditions also provide a means by which to diagnose unsuccessful cases of implicature.[39]

---

[37] The audience might take the speaker to have conversationally implicated Z, when the speaker S intended to implicate Q. In this case, S did implicate Q, but did not *successfully* conversationally implicate anything, either Q or Z.

[38] These success conditions should make it clear that the Conversational Principle (and its related maxims) are not extensional, but dispositional. They do not describe how speakers actually think when generating meaningful utterances. Rather, they are *ex post facto* laws which audiences can use to try to re-construct the speaker's meaning. (Though again, the audience need not, and typically does not, think in these terms, since the mental processes are largely automatic; but they can explicitly reason in this way to the same result.) Speakers can, if they want, explicitly refer to these principles and maxims before speaking as an aid in speaking clearly (though they typically do not). The dispositional nature of the Conversational Principles and maxims can serve as a good example about how to think of other principles and maxims in our theory of moral psychology.

Admittedly, the Determinacy condition denies Grice's closing comment in "Logic and Conversation":

> Since, to calculate a conversational implicature is to calculate what has to be supposed in order to preserve the supposition that the Cooperative Principle is being observed, and since there may be various possible specific explanations, a list of which may be open, the conversational implicatum in such cases will be disjunction of such specific explanations; and if the list of these is open, the implicatum will have just the kind of indeterminacy that many actual implicata do in fact seem to possess. (1975, pg. 58)

Grice's countenance of indeterminacy has elicited considerable comment.[40] Martinich (1984, pg. 511) thought indeterminate conversational implicature accounted for metaphors. He asserts that a poet who states, "My love is a red rose" has conversationally implicated that his love is "beautiful, or sweet-smelling, or highly valued…." Sadock, however, objects, arguing that a metaphor that is not apt is no metaphor at all (1978, pg. 368). Davis (1998, pg. 71–2) rightly objects that the poet does not have just a vague notion in mind. He is not claiming that his love has at least one but not necessarily all of the properties in the open disjunction, and the poet would likely be offended with anyone who says his love smells nice, but is ugly and worthless. My necessary conditions for successful conversational implicature also reject any indeterminate implicature as successful, unless the speaker's meaning was intentionally vague. Otherwise, the speaker has simply failed to convey

---

[39] Success for implicatures is not a binary state, but comes in degrees. The success of an implicature is based on how much of the speaker's intended meaning is conveyed accurately or how closely what H takes to be the meaning matches what S intended.

[40] See also Kempson (1975, pg. 144, 159), Leech, (1983, pg. 23), Sperber and Wilson (1987, pg. 705–6), and Fasold (1990, pg. 132).

his meaning in any clear fashion. He could have and should have done a better job at getting his meaning across.

Returning once again to the Gricean letter, we can now diagnose precisely why Prof. A's implicature was unsuccessful. There were multiple cooperative interpretations of A's utterance that we found, such that we could not decide which among them A was actually implicating. Prof. A did not implicate the disjunct that X was a bad philosopher or a bad instructor. He had a definite intention in mind, to convey that X was a bad philosopher. He just did a poor job of accomplishing that goal. Hence, the Determinacy condition was not satisfied. The likely reason for the multiple interpretations was that there was not sufficient common knowledge between Prof. A and his audience to rule out all but one of the possible interpretations of the utterance. The lack of a clear conversational implicature in this case is not a failure of the theory (which can predict and explain the misunderstanding), but of the speaker, Prof. A, who should have done a better job of making his meaning clear.

Sadock's example of Mary's letter of recommendation makes the problems with indeterminate implicature even clearer. Recall that it is not clear from the letter whether Mary is damning with faint praise or using litotes to understate her praise for Jane. Mary might have meant that Jane is good at philosophy or that Jane is bad at philosophy. But this is not the same as meaning that Jane is either good or bad at philosophy. An indeterminate implicature is not the same as implicating an indeterminate disjunct. The intentions differ radically between them. Just like Prof. A, Mary had a clear intention, but failed to communicate it. She did not have an indeterminate intention and succeed in conveying that. Grice was wrong about the indeterminacy

of implicature. Appealing to success conditions for implicating, however, can easily modify the theory.

4 Jokes, Misleading, and Implicatures

Finally, I would like to close this chapter with an example of the work that my newly expanded and normative theory of implicature can be put. We can consider what is morally and semantically different between deceiving by implicature (misleading)[41] and joking by implicature. In both cases, the regular use of implicature by speakers is being exploited to non-cooperative ends. To understand what I mean, let's take a few example conversations. The first is one of Grice's own examples.

Conversation 4:

A: I am out of petrol.
B: There is a gas station around the corner.[42]

In Grice's original version, *B* has successfully (and truthfully) implicated that petrol is available at the garage. However, let's suppose now that the garage is not open. *B* has not *said* anything false. He did not explicitly state that the garage was open. Nevertheless, *A* takes *B* to have meant that

[41] Deception by implicature – where the speakers intends for the audience to come to believe something false by saying (or making as if to say) something true or without truth value – is generally thought to be distinct from lying. Cf. Addler (1997) and Green (2001). There is supposed to be a parallel between lying/misleading and saying/implicating.

[42] Interesting to note in this example, is the implicature that Grice did not comment upon. A's utterance is not a question, but one is clearly implicated. A has observed the maxim of relation, and so has implicated the question of where he might find more petrol; otherwise, A would be uttering random nonsense with no apparent point.

petrol was available (because the garage is open), and *B* intended for *A* to think this. *B* has deceived to *A* without saying anything false; he's mislead *A*.

Jokes can also exploit the use of implicature. Comedians will use the fact that people will normally take one proposition to be implicated and then humorously present an alternate interpretation. For instance consider these three conversations:

Conversation 5:

C: How was your hunting trip?
D: Pretty good. One morning I shot an elephant in my pajamas. How he got in my pajamas, I don't know.

Conversation 6:

E: Doctor, is it normal to smoke after sex?
F: It's a fairly common behavior.
E: Oh, I figured if I was smoking then we were doing it too fast.

Conversation 7:

G: We have a strict drug-free workplace. Do you do drugs?
H: I used to do drugs.
G: That's okay, so long as you don't anymore.
H: Well, I used to do drugs. I still do, but I used to as well.

The joker in each of these conversations is exploiting the fact that the audience will interpret the utterances in one way, though another is possible. *Conversation 5* (with credit to the great Groucho Marx) involves exploiting the syntax rules of word order for prepositional phrases. *Conversation 6* doesn't involve an implicature, but rather the pragmatic filling in of what is said due to the lexical ambiguity of 'smoke.' *Conversation 7* plays on the non-conventional implicature from *H*'s first utterance. The fact that *H* initially made no reference to his current usage leads *G* to take *H* to mean that he no longer uses drugs, on the assumption that *H* is being cooperative and observing the Maxims of Relation and Quantity.

Let's stick with *Conversation 7* for now, as it is the closest parallel with *Conversation 4*. In both cases, the speaker is not being conversationally cooperative: he is not going along with the purpose or direction of the talk exchange, which was set by his dialogue partner. In the case of this joke, a joke was not asked for; it wasn't part of the set direction of the dialogue. Even if G ended up thinking it was funny, it still wasn't conversationally cooperative. It would seem the purpose or direction of the talk exchange can't be retroactively defined to make an unexpected joke cooperative. What about communicative cooperation? The liar is clearly being uncooperative in that regard. The joker's utterance is more complex. His initial utterance – the setup for the joke – was meant to be misunderstood. But he also intended for his meaning to become clear, not just in regards to his drug usage, but also that he was telling a joke. A joke must be recognized as such to be successful, much like an implicature must be understood to be successful.

These considerations suggest that an initial answer to our question about the difference between *Conversations 4* and *7*: the distinction lies in whether or not the deception was intended to be revealed. The joker intends for his utterance to be recognized as a joke, eventually. The liar never does. So jokes involve a delayed communicative cooperation, while misleading does not. While there is something correct there, this answer misses the mark. There is every reason to suggest that *B* expects *A* to go around the corner looking for petrol, discover the closed garage, and realize he's been mislead. So the deception will be revealed too. While it isn't the case that all (or even most) cases of misleading involve the expectation by the speaker that the deception will be revealed at some point, we cannot rule out cases like *Conversation 4* by fiat from counting as genuine examples of misleading.

Another possible answer then is that the real difference resides in the fact that *H* expects (or hopes) G will appreciate the joke, while *B* has no such expectations of *A*. We are then brought back to the notion of retroactively defining the purpose of the talk exchange. G wasn't expecting a joke, but retroactively agreed to it, so that *H* ended up being conversationally cooperative, though *B* did not. This is not to say that the purpose was retroactively modified; the joke remains non-conversationally cooperative. But *H* expects or hopes that G will appreciate the joke anyway. *H* intends it to be amusing to his audience while *B* does not.

This answer again has potential, but problems remain. As a persistent punster myself, I know all too well that jokes are not always appreciated by my audiences. But that has nothing to do with whether or not they *get it*. Sometimes speakers tell jokes for their own amusement without expecting the audience to appreciate them. In *Conversation IV*, G might be angered by the joke because he takes this to be a very serious matter.

All this analysis goes to show the complexity involved in separating joking by implicature from misleading. Perhaps they can't be definitively separated. The point of this brief foray into the question is meant to demonstrate the work that virtue-theoretic semantics can be put to. In what way are these two conversations different both in terms of whether they were morally permissible and how the speakers came to communicate what they did? And can the answer to one of those questions will help answer the other. The close connection for which I've been arguing between meaning and morality suggests that the two questions are closely linked and to answer one requires answering the other.

Chapter 7: Normative Implicature and the Is-Ought Problem

The question how moral statements can be proven is thought to be critical for ethics. Moral claims are supposed to be wholly distinct from statements of fact. Consequently, non-naturalists assert that normative statements cannot be inferred from factual ones. This is the long-standing is-ought problem, widely thought to challenge much of moral philosophy. Various attempts to defuse the problem have met with limited success. By making use of the normative nature of implicature and cooperation that we've now established, I argue that the argument for the is-ought problem contains a clear, but heretofore unrecognized, implicature that commits the is-ought fallacy. Stating the is-ought problem implicates that one ought not derive an 'ought' from an 'is.'

1 The Is-Ought Argument

Normative statements differ from descriptive statements in the sense that that the former cannot be derived from the latter. Many have thought this to be a damning objection against much of naturalized ethics. Hume (1739: 469) appears to have presented the problem in *A Treatise on Human Nature*, though the interpretation is debated.[1] From him, the problem took on the appellation 'Hume's Guillotine.' Karl Popper (1948) articulated a modern version.

We can state the problem this way: no evaluative or normative statements can be validly inferred from a set consisting of only descriptive statements. Since the introduction of the is-ought problem, Searle (1964) offered counterexamples, Black (1964) tried to close the gap between 'ought' and 'is,' and Singer (1973) argued the problem is trivial. Most non-naturalists, however, have been left unconvinced, maintaining that moral facts are fundamentally distinct from natural facts. In what follows, I present a new approach to dealing with the is-ought problem. Instead of working to establish some way to validate normative inferences from factual premises (as has been the most prevalent strategy to overcome the problem), I contend by arguing against the is-ought fallacy, one commits the fallacy.

That there is an is-ought fallacy is sometimes thought to be obvious. For clarity, let's construct the argument establishing this conclusion, which we can dub the is-ought argument.

---

[1] I am not here primarily concerned with Hume's discussion, but instead focus on a modern version of the problem. If in fact Hume did express something akin to the modern Is/ought problem (which I take him to have done), then he will have fallen victim to the fallacy himself. He recommends that his readers notice the move in the writings of ethicists from 'is' to 'ought'. That recommendation, if not implicitly containing an 'ought' (they ought to look for it), seems to clearly implicate it.

The argument, as I see it, has three premises (**P1**, **P2**, **P3**), a stated conclusion (**SC**), an entailed conclusion (**EC**), and an implicated conclusion (**IC**). The argument proceeds as follows:

P1 The class of evaluative statements is not identical to the class of factual/descriptive statements.[2]

P2 Statements of one class can be validly inferred from another class of statements only if the inferred class is identical to or supervenient upon the other class.

P3 Evaluative statements do not supervene upon factual/descriptive statements.

Therefore:

SC Evaluative statements cannot be validly inferred from factual/descriptive statements.

EC It is a fallacy to infer evaluative statements from factual descriptive statements.[3]

IC One ought not commit this fallacy.[4]

Roughly, factual/descriptive statements are those saying what is the case. They describe the facts of the natural world, and their truth conditions are based upon the natural world.[5] By contrast, evaluative statements state what ought to be the case.[6] They establish or appeal to norms.

---

[2] Searle (1964), formulating the problem similarly, referred to evaluative statements instead of (moral) 'ought' statements as Hume (perhaps), Popper, et al. discussed. I discuss possible differences between evaluative statements and moral statements later. Evaluative statements at least encompass all moral, ought statements.

[3] I will consider in section III the possibility of an additional, unstated premise.

[4] **IC** need not take precisely this form. Someone articulating the argument could just as well implicate something like "This fallacy ought to be avoided" or "Everyone ought to avoid making this fallacy." What precisely is implicated does not matter greatly for our purposes.

[5] I am being somewhat loose here, for the statements themselves do not mean or do these things, but rather the speakers uttering them on a particular occasion. But for present purposes we may abstract to what speakers typically mean when they utter statements of either type.

[6] One might contend that these two categories are not mutually exclusive. Some statements could express the fact that there is a norm which ought be followed; such statements might seem both factual and evaluative. However, this view inherently undermines the factual/evaluative distinction upon which the entire is-ought problem has always been based, and the proponent of the is-ought fallacy cannot appeal to it.

They evaluate (roughly) whether something is good or bad, appropriate or not, desirable or not, acceptable or not.[7] All moral statements are evaluative, but not all evaluative statements are moral. "I ought to be a millionaire" is an evaluative statement, though perhaps not a moral one. It expresses a judgment about the desirability (for me) of a certain state of affairs. It is distinct from a factual claim about how the world is, such as "I desire to be a millionaire." This distinction is widely accepted in the literature, and most challenges to the is-ought argument do not challenge **P1** or **P3**.

Instead, objections typically focus on **P2**, often by offering purported counterexamples. However, for the sake of argument, I accept (and take it that non-naturalists accept) that **SC** is a valid conclusion from **P1**-**P3**. That **SC** entails **EC** seems obvious. If a type of inference is not valid, then making that inference is committing a fallacy. So the is-ought fallacy is established.[8]

All of naturalized moral philosophy is supposed to be guilty of committing the is-ought fallacy. The trouble for the is-ought argument, however, comes in the implicated conclusion that one ought to avoid committing this fallacy. The implicature is an evaluative statement, yet it is implicated by and to be believed on the basis of exclusively factual/descriptive statements.

---

[7] We should be mindful that some evaluative statements are cloaked in factual/descriptive form, so that they look to be describing the world, but are still appealing to some norm. For instance, Searle gives the example "John is under obligation to pay Smith five dollars" (1964, pg. 44). He claimed this is a factual/descriptive statement, and then used it to generate an evaluative inference that John ought to pay. I take it that the notion of obligation is already a loaded evaluative term, so that it is a disguised evaluative statement since it expresses a norm. His conclusion that John ought to pay his debt then does follow, but it no longer is a counterexample to the is-ought fallacy.

[8] Though related to G. E. Moore's naturalistic fallacy, the is-ought fallacy is distinct from it. To be clear, we should note that if an evaluative statement is included among the premises, then one could validly infer a subsequent evaluative statement.

2 The Is-Ought Implicature

The idea behind claiming that the implicated conclusion **IC** is present is that the whole point of stating the is-ought argument is to inform philosophers about what kind of arguments they ought not make. To simply note that a fallacy exists, while true, is by itself uninteresting. The larger point behind stating that such-and-such is a fallacy is to demonstrate that a particular logical move is inappropriate; it *should not* be used. But this intent alone is insufficient to establish the presence of an implicature.

Paul Grice developed the notion of implicature in "Logic and Conversation" (1975). An implicature is something meant or suggested distinct from what is literally said, which the speaker intends for their audience to recognize as meant or suggested. For instance, in response to *A*'s remark of "I'm thirsty," *B* replies, "There's beer in the fridge," implying that *A* is welcome to drink one.[9] An implicature is always cancellable without contradiction. *B* can then state, "But you can't have one." Additionally, an implicature can be distinct from what an utterance directly entails. *B* has not implicated the entailment "There's a brownish, alcoholic liquid in the fridge."[10]

Someone articulating the is-ought argument can generally be taken to conversationally implicate **IC**. This case conforms to Grice's first example of generating conversational implicature, where the presence of an implicature prevents the violation of any type of conversational maxim (Quality, Quantity, Relation, and Manner). Whether or not one states the entire argument or

---

[9] *B* has also likely implicated that beer will quench *A*'s thirst, and possibly that *B*'s expectation that *A* will drink a beer if offered one.

[10] I am not claiming that **IC** is entailed, but implicated; hence this is not a case of question begging against the is-ought argument. The problem arises because if the speaker of **SC** does implicate **IC**, then the speaker does take it that **IC** follows from **SC**, though he just said that is not a valid move.

simply **SC** is irrelevant to the generation of an implicature, so let us simplify matters by imagining the following conversation between John Stuart Mill and David Hume.

> **John**: There is untold suffering in Africa due to famine, so I ought to donate money to the relief efforts.
>
> **David**: Evaluative statements cannot be validly inferred from factual/descriptive statements.

Initially, without supposing that David has implicated anything, David's response does not appear to be a cooperative contribution to the conversation. The semantic content of what he said is by itself not immediately germane. If, however, David is being cooperative (and so observing the maxim of Relation – *be relevant*), then it is reasonable to infer that David has implicated first that John has committed this fallacy and second that he ought not to have done so. The first implicature is factual and uncontroversial. It is not alone sufficient, however, to make the remark relevant and cooperative. What is the point of simply demonstrating that John has committed a fallacy without meaning that it is *wrong* to do so? David means to correct John's mistake. The same holds true when a professor explains to an undergraduate that his conclusion does not follow because of an *post hoc ergo propter hoc* fallacy. The professor is not merely telling the student about another type of fallacy. She means that the student should not make this mistake; she has evaluated his argument. Likewise with David. He has evaluated John's statement and found it wanting. We can take it that David thinks it is true that John ought not to have committed this fallacy. He means to convey his evaluation of John's statement by his utterance, and he means to give John a reason to think his evaluation is correct. David believes the legitimacy of his implicated evaluation derives from the truth of his utterance, which is a factual/descriptive statement.

The same line of reasoning applies to the is-ought argument. *Ex hypothesi*, **IC** is not entailed by **SC**, **EC**, or **P1–P3**. However, it is typically[11] the case that one articulating the is-ought argument intends to implicate **IC** in order to explain what one ought not do. Imagine that David articulates the is-ought argument to John, stating **SC**. David believes **IC** to be true; one ought not to commit the is-ought fallacy. David also wants to give his audience a reason to believe **IC** is true. The validity of **IC** will either come from **SC** (a factual/descriptive statement) or from some other unstated premises. Since David has not provided any other reason to support **IC**, it is reasonable to conclude that David regards **SC** as establishing **IC**. But David is also committed to the truth of **SC**, which entails that **SC** cannot establish the truth of **IC**. He has run afoul of the very fallacy he sought to damn. By condemning the is-ought fallacy, you commit the is-ought fallacy.

As it turns out, Hume's Guillotine is appropriately named. One could use it to induce a reign of terror in order to free philosophy from the tyranny of naturalized ethics. But, like Robespierre, those revolutionaries who use Hume' Guillotine tend to be fall under its blade, thereby ending the terror. Employing Hume's Guillotine to oppose the non-naturalized revolution in ethics is itself counter-revolutionary and unsustainable.

3 Objections

One might object that there is an unstated premise at work, which would state that one should not make fallacious arguments. This is the **Assumed Premise Objection** and represents

---

[11] The 'typically' proviso is present because a speaker can cancel the implicature (though the speaker is then not being cooperative and relevant) or the conversation may simply be about types of fallacies, in which case the speaker might not implicate (and mean) anything evaluative about the inappropriateness of committing this fallacy.

the most serious attack upon my argument. Suppose there is an assumed premise **AP**: "If an argument is fallacious, then one should not make it." **AP** is an evaluative statement. If **AP** is added the other premises, then we no longer have a set of only factual/descriptive statements, as we did before. The is-ought argument does not object to deriving an evaluative statement from a set containing both evaluative and factual statements. Hence, with the addition of **AP**, **IC** can now be derived without any problem. So the Assumed Premise Objection claims. I admit that adding **AP** can prevent David from committing the is-ought fallacy. It is the only escape route to this problem of which I am aware. However, there are two responses to this objection. First, it is not available to all meta-ethical theories. Second, the supposition that there might be an assumed premise is far short of an argument that there is an assumed premise.

Not everyone can appeal to the Assumed Premise Objection. The objection grants that an argument includes evaluative statements, and so must express propositions. Hence, all forms of non-cognitivism are barred from raising the assumed premise objection. Both **AP** and **IC** are taken to be true evaluative statements, and so neither error theory nor any type of moral anti-realism can use this objection. The Assumed Premise Objection grants that there is a fallacy that ought to be avoided, and fallacies are not subjective. Therefore, moral subjectivism is also ruled out.

Any of remaining theory must also have at least some evaluative statements that are primitive, requiring no derivation of their truth. The question is how do we know **AP** is true? If **AP** is primitive, then we just know it. If **AP** is not primitive, then it must be derived from other set of premises which must include at least one evaluative statement(s). Thus, any non-naturalist theory without at least one primitive evaluative statement will be lost in an infinite regress: **AP** is

derived from another evaluative statement, which itself is derived from a third evaluative statement, requiring a fourth, and so on without end.

There are a few versions of non-naturalism that remain capable of appealing to the Assumed Premise Objection. Ross's ethical intuitionism (1930) serves as an example, in which some moral statements are *prima facie* true, requiring no further proof. For such theories, **AP** may be a primitive evaluative statement, obviously true and not in need of any derivation.[12] **IC** may then be derived from **P1**, **P2**, **P3**, and **AP**.[13] Though views like Ross's can refer to the Assumed Premise Objection, the presence of an unstated premise has thus far gone unrecognized and un-argued in the literature. Only a limited number of theories can take this one escape route. Yet, before they do, the actual presence of an assumed premise must be argued for. It is not enough that the **AP** might be there.[14]

There are a few other objections possible against this view of a problematic implicature, all of which fall short. First, suppose someone disputes the existence of an implicature. On this view, David did not implicate that John was wrong (an evaluative notion) or that John ought not to commit this fallacy. Rather, everything David meant is entirely captured by what he said, and that is completely absent of evaluation. I am not suggesting that there are no conditions under which someone uttering **SC** would not implicate **IC**. There surely can be a few cases. For instance, if two

---

[12] Alternatively, **AP** may be derivable from some other primitive evaluative statement.

[13] Note that this argument does not deny that **IC** is true and part of the intended meaning of uttering the is-ought argument.

[14] The presence of **AP** will be a problem for the relevance theory of Sperber & Wilson (2002), since a Gricean account does not need **AP** to explain how **IC** is implicated, but they will require it to show relevance.

logicians are listing fallacies and one utters **SC**, it is possible that **IC** would not be implicated. However, in conversations such as that between John and David, David's response can only be seen as relevant if the implicature is present.

A second objection grants that **IC** is validly inferred, but denies that it was inferred from solely factual/descriptive statements. Rather, it argues that entailed conclusion **EC** is a disguised evaluative statement, because the intuitive notion of a fallacy is inherently normative. One then infers **IC** from **EC**. The is-ought argument does not forbid inferring an evaluative statement from a set containing both factual/descriptive and evaluative statements. So, the implicature is not fallaciously derived. This would be a fine objection but for the fact that it simply changes where the fallacy is committed. The objection grants that **SC** entails **EC**. But now **EC** is an evaluative statement. So, **EC** entailed solely by factual/descriptive claims. Hence, the is-ought fallacy is still committed, and there was never any need to bring up **IC**.[15]

A third objector might take a slightly different tack by claiming that **SC** is an illocutionary act. Thus, instead of an implicature, **SC** carries an illocutionary force urging John not to reason fallaciously. J. L. Austin (1975) first developed the notion of illocutionary acts to explicate cases in which the utterance of a statement performs an action, so what is meant exceeds the propositional content expressed. An example case is someone saying, "I take this woman to be my lawfully wedded wife." The speaker has performed an action beyond stating what he is doing: he actually does something (marries the woman in question) by saying these words. The argument here then is that David's utterance of **SC** has the illocutionary force of urging John to avoid this fallacy

---

[15] Claiming instead that **P2** is a disguised evaluative statement ends with much the same result.

without actually implicating that he ought to do so. The problem with this argument is that by Austin's own account, illocutionary acts can still imply something else. For instance, in the example above, Austin admits that the groom has implied that he is not already married. Therefore, even if we were to regard David's utterance of **SC** as an illocutionary act urging John to avoid the fallacy, this illocutionary force does not prevent David from also implicating **IC**. In fact the implicature gives the reason behind the illocutionary force.[16]

A fourth objection makes a related argument, claiming that **IC** is not implicated, but entailed. This cannot be so. Recall that implicatures can be cancelled; entailments cannot. So long as it's not stated, **IC** can be cancelled. David could continue on to say, "But I don't mean that you ought not commit this fallacy," in which case he did not implicate **IC**. (Cancelling **IC** in this way, however, would be quite odd and leave John wonder as to the point of David's initial utterance.) Additionally, were this objection correct, then it would be of little assistance to anyone favoring the is-ought argument. The objection concedes that the fallacy is committed. **IC** is clearly an evaluative statement, and the argument admitts that it was validly inferred exclusively from factual/descriptive statements. This fourth objection is doomed from the outset and may be set aside.

The fifth and final objection contends that there are two types of evaluative statements (at least): moral and logical. And the 'ought' in **IC** is not the same sort of 'ought' with which the argument takes issue. The is-ought argument was meant to object to ethics and the common fallacious reasoning to moral statements. What we may call 'the logical variety of ought' refers to

---

[16] I am skeptical of **SC** being an illocutionary act, and here only treat it as such for the sake of argument.

statements condoning or condemning certain forms of inference. To claim, for instance, that one ought not deny *modus ponens*, though evaluative, is not a moral claim. Likewise **IC** contains a logical ought, not a moral one. The is-ought fallacy, so the objection goes, only pertains to moral ought's. Hence, the fallacy is not committed.

For a moment, let's grant the notion that there are these two types of ought, so that to say one ought only use valid arguments is not a moral demand. However, both kinds of 'ought' are both evaluative principles. This objection does not claim that logical ought statements are a special sort of evaluative statement that (unlike the other evaluative statements) are either identical to or supervenient upon factual/descriptive statements. Logical ought statements remain a distinct from factual/descriptive statements. So how can they be derived from factual/descriptive statements? What this objection needs (but fails to provide) is a principled reason to explain why logical ought statements can be validly derived from solely factual/descriptive statements, but other sorts of evaluative statements (such as moral claims) cannot be so derived. Lacking such a principle, it must be concluded that while **IC** may not be a moral ought statement, it is still an evaluative statement. And no evaluative statement of any kind can be validly derived from factual/descriptive statements.

4 Conclusion

I have not claimed there is anything wrong with the is-ought argument. It validly concludes that there is a fallacy. Making the argument, however, carries an evaluative implicature, one derived solely from factual statements. The implicature is unavoidable. It might be that this implicature is unproblematic, due to an unstated, normative premise, from which the normative

implicature legitimately follows. No one has ever claimed that the is-ought argument contains this premise. Additionally, not all proponents of the fallacy can make use of it. More likely instead is the conclusion that due to the implicature, denouncing the is-ought fallacy commits the fallacy.

Chapter 8: Conclusions

The central claim that I have been advocating is that we simply cannot separate value and normativity from meaning and language use. Our ability to use language meaningfully derives in large part from our status as moral agents. I began by establishing an outline of Rational Virtue Theory, a novel, quasi-Gricean theory of moral psychology. The basic tenets of this view are that (1) we are rational, egoistic creatures who seek *eudaimonia*; (2) achieving *eudaimonia* means realizing enough of our ends, our life goals, which in turn means heeding certain End Selection Rules and End Balancing Rules; and (3) there are certain patterns of behavior we should adopt – the Behavioral Principles – that help us to realize our ends, and they are widely applicable to other rational egoistic creatures like us. Because the End Rules and Behavioral Principles are the same for everyone (since everyone can separately find the rules and principles to be helpful guidelines in their own individual quests for *eudaimonia*) we can compose a general manual for rational, ethical behavior.

The manual is far from complete, but I have provided the beginnings of an account of how to go about developing it. We know what its contents need to look like. And there are three main sources to turn to for the content. First, we can look to empirical research on happiness to gain a further idea of what human flourishing means. Perhaps, as Grice suggested, there is some feature that unifies it. Or maybe the ways to be happy are so diverse that there is no unifying feature. The key point is that we decide for ourselves what *eudaimonia* consists in. We do at least know that it is a complex end. To be happy we need to realize a number of ends that are desirable for their own sakes and desirable for the sake of happiness. Even with a great diversity of views of what those ends are (or should be), there are certain rules for how to select and balance all of our ends that apply to everyone. We can choose different things, but the criteria for choosing are the same for us all. Conceptual analysis will be the greatest tool in exploring what the End Selection and Balancing Rules are. When it comes to how to behave, we can look to game theory and decision theory. Whatever our ends are, there are certain ways of behaving that are the most conducive to realizing them, especially since we live in an environment surrounded by creatures trying to do the same thing. We have already found cooperation to be one such pattern of behavior. More should follow.

The moral psychology I have developed underpins the rational, cooperative nature of communicative language use. By establishing the normative nature of cooperation, particularly in language use, I have explained why Grice's own account of implicature needs revision. First, I distinguished between Conversational Cooperation and Communicative Cooperation, with the latter being the actual source of the conversational maxims instead of the former, as Grice assumed. Second, I argued that moral maxims can and do play the same role as the conversational

maxims in generation and interpretation of non-conventional implicature. This realization then led to a new taxonomy of implicatures. It also indicated that cases of implicature through maxim flouting are rare. Instead, much more common are cases of maxims clashing, where one of the maxims is something other than Grice's original list of conversational maxims. Third, through a new analysis of the Gricean letter of recommendation case, I found that there is a new category of Conversational Principles – of which the Cooperative Principle is but one – that play a role in explaining the content of an implicature.

These results speak to the need for a virtue-theoretic semantics, according to which communication is a normative activity. Likewise then the study of meaning and communication must be understood as a normative discipline. The discussion of Communicative Cooperation and of successful implicature that I presented should serve as a basis for this emerging virtue-theoretic semantics. The main idea of this theory will be that there are certain habituated patterns of behavior that are conducive to meaningful communication. Virtuous utterers are those who are more regularly successful in making their meanings understood. Virtuous audiences are those who more regularly understand correctly an utterer's intended meaning. We should be able to identify various virtues that these agents regularly display. Cooperation is already one such virtue that we identified. Some discussion has been given to politeness, but as I argued, much of it misses the mark. More specific sub-maxims are needed than just "be as polite as is required" in order to be sufficiently informative to our theory. The investigation into virtue-theoretic semantics will ask what are the values that regularly and consistently affect the way that we speak or interpret others and how do they have that effect.

The study of virtue-theoretic semantics can proceed on a couple of fronts. First, I have already discussed the relationship between misleading and joking by implicature that warrants further exploration. These are only two types of cases of implicature in morally charged communication that are now open to examination.

Second, empirical research can be done looking for various communicative norms and their roles in generating and interpreting meanings. For instance, we might suppose that being polite in a particular way is a communicative norm. To test this, we can examine how audiences interpret differently an utterer's utterance when that norm has been violated versus when it has not been violated. The addition of an insult could be an example, as in chapter 6. Will speakers interpret "There is a garage around the corner" differently than "There is a garage around the corner, moron?"

In the end, the key point is that our communication is ethically charged and normatively governed. We adhere to moral maxims from our moral psychology when we speak. And audiences expect us to since they come to understand our meaning on the basis of that assumption. Though Grice appears to have had some notion of normativity in his theory of implicature, I have come in to expand and clean up that theory. For Grice, implicature helps fill in the gap between what is saying and meaning. By examining Grice's assumed notions of rationality and cooperation, I have discovered new types of implicature and expanded how much of that gap implicature fills in. In the process, I have argued for the moral component of language use – which Sperber and Wilson (1986) deny as mere brute facts of human psychology – which are rational and in our self-interest to follow. By understanding that fact, we better understand how and why we use language.

Bibliography

Ackrill, J. L. (1974). *Aristotle on Eudaemonia*. Oxford: Oxford University Press.

Adler, J. (1997). Lying, Deceiving, or Falsely Implicating. *Journal of Philosophy* 94. 435–452

Alexander, J. M., & Skyrms, B. (1999). Bargaining with Neighbors: Is Justice Contagious?" *Journal of Philosophy*, 96: 588–98.

Annas, J. (2006). Virtue Ethics. In David Copp (ed.), *The Oxford Handbook of Ethical Theory*. (515–536). Oxford: Oxford University Press.

Aristotle. (1925). *Nicomachean Ethics*. (D. Ross, Trans., J. L. Ackrill and J. O. Urmson, Revised). Oxford: Oxford University Press.

Arrow, K.J. (1965) The theory of risk aversion, in *Aspects of the Theory of Risk Bearing*, by Yrjo Jahnssonin Saatio, Helsinki. Reprinted in: *Essays in the Theory of Risk Bearing*, Chicago: Markham Publ. Co, 1971, 90–109.

Atlas, J. D. (1979). How linguistics matters to philosophy: Presupposition, truth, and meaning, in C.-K. Oh, & D. A. Dinneen (eds.), *Syntax and Semantics, 11: Presupposition*, pp. 265–81. New York: Academic Press.

Austin, J. L. (1975). *How to Do Things with Words*. Cambridge, Mass.: Harvard University Press.

Bach (2006). The Top 10 Misconceptions about Implicature. In B. J. Briner and G. Ward (Eds.), *Drawing the Boundaries of Meaning*. (21–30). Philadelphia: John Benjamins Publishing.

Bach, K. (2004). Pragmatics and Philosophy of Language. In L. R. Horn and G. Ward (Eds.), *Handbook of Pragmatics*. (463–487). Malden, Massachusetts: Blackwell.

Bach, K. (1987). On Communicative Intentions: A Reply to Récananti. *Mind and Language*, 2, 141–154.

Baker, J. (1991). Introduction. In P. Grice, *The Conception of Value*. (J. Baker, Ed.). (1–21). Oxford: Clarendon Press.

Bach, K., and Harnish, R. (1979). *Linguistic Communication and Speech Acts*, Cambridge, MA: MIT Press.

Baker, J. (1989). The Metaphysical Construction of Value. *The Journal of Philosophy*, 86. 503–513.

Baker, J. (1986). Do One's Motives Have to be Pure? In R. E. Grandy and R. Warner (Eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. (457–474). Oxford: Clarendon Press.

Barnes, G. (1993). Review [Review of the book *The Conception of Value*]. *Mind*, 102, 366–370.

Bennett, J. (1976). *Linguistic Behavior*. Cambridge: Cambridge University Press.

Berg, J. (1991). The relevant relevance, *Journal of Pragmatics*, 16: 411–25.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Bicchieri, C. (2000). Words and Deeds: A Focus Theory of Norms. In J. Nida-Rumelin & W. Spohn (Eds.), *Rationality, Rules, and Structure*. Dordecht: Kluwer Academic Publishers.

Bicchieri, C. (1997). Learning to Cooperate. In C. Bicchieri, R. Jeffrey & B. Skyrms (Eds.), *In The Dynamics of Norms*. Cambridge: Cambridge University Press.

Bicchieri, C. (1990). Norms of Cooperation. *Ehics*, 100, 838–861.

Black, M. (1964). The Gap Between 'Is' and 'Should'. *The Philosophical Review*, 73, 165–181.

Blakemore, D. (1992). *Understanding Utterances*, Oxford: Basil Blackwell.

Blanchflower, D. & Oswald, A. (2007). Is Well-being U-Shaped over the Life Cycle? *The Warwick Economics Research Paper Series* (TWERPS), 826.

Brown, P., and Levinson, S. C. (1978). Universals in language use: Politeness phenomena, in E. Goody (ed.), *Questions and Politeness*, pp. 56–311, Cambridge: Cambridge University Press.

Camerer, C. (2003). *Behavioral Game Theory: Experiments on a Strategic Interaction*. Princeton, NJ: Princeton University Press.

Campbell. A. (1981). *The sense of well-being in America*. New York: McGraw·Hill.

Carnap, R. (1937). *The Logical Syntax of Language*. (Amethe Smeaton, Trans.). London: K. Paul Trench.

Carrol, L. (1896). *Symbolic Logic, Part I*. London: Macmillan and Co.

Carrol, L. (1871). *Through the Looking Glass, and What Alice Found There*. London: Macmillan and Co.

Clark, A. E., Oswald, A., Warr, P. (1996). Is job satisfaction U-shaped in age? *Journal of Occupational and Organizational Psychology*, 69. 57–81.

Costa, P.T., Jr., & McCrae. R.R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology*, 38, 668–678.

Crohan, S.E., Antonucci. T.C., Adelmann, P.K., & Coleman, L.M. (1989). Job characteristics and well-being at midlife. *Psychology of Women Quarterly*, 13. 223–235.

Davis, W. A. (1998). *Implicature: Intention, Convention, and Principle in the Failure of Gricean Theory*, Cambridge: Cambridge University Press.

Dernber, W.N., & Brooks, J. (1989). A new instrument for measuring optimism and pessimism: Test-retest reliability and relations with happiness and religious commitment. *Bulletin of the Psychonomic Society*, 27.365–366.

Diener, E. & Biswas-Diener, R. (2002). Will Money Increase Subjective Well-Being? *Social Indicators Research*, 57. 119–169.

Diener, E. & Diener, C. (1995). The Wealth of Nations Revisited: Income and Quality of Life. *Social Indicators Research*, 36. 275–286.

Diener, E., Lucas, S. Oishi and E. M. Suh (2002). Looking up and looking down: Weighting good and bad information in life satisfaction judgments' *Personality and Social Psychology Bulletin*, 28. 437–445.

Diener, E., Sandvik, E., & Larsen, R.J. (1985). Age and sex effects for emotional intensity. *Developmental Psychology*. 21, 542–548.

Doris, J. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Emmons, R.A., & Diener, E. (1986). Influence of impulsivity and sociability on subjective well-being. *Journal of Personality and Social Psychology*. 50. 1211–1215.

Emmons, R.A., & Diener, E. (1986a). An interactional approach to the study of personality and emotion. *Journal of Personality*. 54,371–384.

Farrell, J., & Rabin, M. (1996). Cheap Talk. *The Journal of Economic Perspectives*, 10, 103–118.

Fasold, Ralph. The Sociolinguistics of Language. Oxford: Basil Blackwell, (1990).

Freedman, J. (1978). *Happy People*. New York: Harcourt, Brace, Jovanovich.

Gazdar, G. (1979). *Pragmatics: Presupposition, Implicature, and Logical Form*. New York: Academic Press.

Grandy, R. & Warner, R. (1986). Paul Grice: A Review of His Work. In R. E. Grandy & R. Warner (Eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. (1–44). Oxford: Clarendon Press.

Green, G. M. (1989). *Pragmatics and Natural Language Understanding*, Hillsdale, N. J.: Erlbaum Associates.

Green, M. S. (1995). Quality, volubility, and some varieties of discourse, *Linguistics and Philosophy*, 18: 83–112.

Green, S. P. (2001). Lying, Misleading, and Falsely Denying: How Moral Concepts Inform the Law of Perjury, Fraud, and False Statements. *Hastings Law Journal*, 53. 157–212.

Grice, P. (2001). *Aspects of Reason*. (R. Warner, Ed.). Oxford: Clarendon Press.

Grice, P. (1991). *The Conception of Value*. (J. Baker, Ed.). Oxford: Clarendon Press.

Grice, P. (1998). Retrospective Epilogue (Strand Six) (1987). In A. Kasher (Ed.), *Pragmatics* (Vol. IV). (177–180). London: Routledge.

Grice, P. (1989). *Studies in the Ways of Words*. Cambridge, Massachusetts: Harvard University Press.

Grice, P. (1986). Reply to Richards. In R. E. Grandy & R. Warner (Eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. (45–106). Oxford: Clarendon Press.

Grice, P. (1975a). Logic and Conversation. In D. Davidson & G. Harmon (Eds.), *The Logic of Grammar*, Encino, California: Dickinson Publishing, 64–75.

Grice, P. (1975b). Method in Philosophical Psychology (From the Banal to the Bizarre), *Proceedings and Addresses of the American Philosophical Association*, 48, 23–53.

Grice, P. (1971). Intention and Uncertainty. *Proceedings of the British Academy*, 57, 263–79.

Grice, P. (1969). Utterer's meaning and intentions. *Philosophical Review*, 78: 147–77.

Grice, P. (1961). The causal theory of perception. *Proceedings of the Aristotelian Society*, Supplementary Volume, 35: 121–52.

Grice, P. (1957). Meaning. *Philosophical Review*, 66: 377–88.

Grice, P., & Baker, J. (1985). Davidson's on 'Weakness of the Will'. In B. Vermazen & M. Hintikka (eds.), *Essays on Davidson: Actions and Events*, Oxford, Clarendon.

Guarini, M. (2000). Horgan and Tienson on Ceteris Paribus Laws. *Philosophy of Science*, 67, 301–315.

Haybron, D. (2009). *The Folk Concept(s) of Happiness: Preliminary Notes*. Retrieved from: http://2600005669687546883-a-1802744773732722657-s-sites.googlegroups.com/site/danhaybron/research/happiness-and-well-being/Thefolkconceptofhappinessnotesv2.pdf

Haybron, D. (2008). *The Pursuit of Unhappiness: The Elusive Psychology of Well-Being*. Oxford: Oxford University Press.

Haybron, D. (2003). What Do We Want from a Theory of Happiness? *Metaphilosophy*, 34, 305–329.

Harman, G. (2000). The Nonexistence of Character Traits. *Proceedings of the Aristotelian Society* 100. 223–226.

Harms, W. & Skyrms, B. (2008). "Evolution of Moral Norms," in *The Oxford Handbook of Philosophy of Biology*, Oxford: Oxford University Press.

Harnish, M. (1976). Logical Form and Implicature. In T. Bevar *et al.* (eds.), *An Integrated Theory of Linguistic Ability* (313-391). New York: Crowell.

Headey, B., & Wearing, A. (1992). *Understanding happiness: A theory of well-being*. Melbourne: Longman Cheshire.

Herzog, A:R., Rogers, W.L., & Woodworth, J. (1982). *Subjective well-being among different age groups*. Ann Arbor: University of Michigan, Survey Research Center.

Hintikka, J. (1986). Logic of Conversation as Logic of Dialogue. In R. E. Grandy & R. Warner (Eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. (259–276). Oxford: Clarendon Press.

Hirschberg, J. (1991). *A Theory of Scalar Implicature*, New York: Garland.

Hugly, P., & Sayward, C. (1979). A Problem About Conversational Implicature. *Linguistics and Philosophy*. 3, 19–25.

Hume, D. (1739). *A Treatise on Human Nature*. Oxford: Oxford University Press.

Inglehart, R. (1990). *Culture shift in advanced industrial society*. Princeton, NJ: Princeton University Press.

Kasher, A. (1998). Conversational Maxims and Rationality. In A. Kasher (Ed.), *Pragmatics* (Vol. IV). (181–198). London: Routledge.

Keenan, E. (1976). The Universality of Conversational Implicature. *Language in Society*. 5, 67–80.

Kraut, R. (1979). Two Conceptions of Happiness. *The Philosophical Review*, 88, 167–197.

Lakoff, R. (1977). What you can do with words: Politeness, pragmatics and performatives, in R. Rogers, R. Wall & J. Murphy (eds.),*Proceedings of the Texas Conference on Performatives, Presuppositions and Implicatures*, pp. 79–106. Arlington, Va.: Center for Applied Linguistics.

Lamberts, K. (1997). Process Models of Categorization. In K. Lamberts & D. Shanks (Eds.), *Knowledge, Concepts, Categories* (371–404). Cambridge, Massachusetts: MIT Press.

Larson, R. (1989). Is feeling "in control" related to happiness in daily life? *Psychological Reports*. 64. 775-784.

Latten, J.J. (1989). Life-course and satisfaction, equal for every-one? *Social Indicators Research*. 21, 599–610.

Layard, R. (2005). *Happiness: Lessons from a New Science*. New York: The Penguin Press.

Leech, G. (1983). *Principles of Pragmatics*. London: Longman.

Levinson, S. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Lewis, D. (1969). *Convention*. Cambridge, Massachusetts: Harvard University Press.

Lewis, D. (1975). Language and Languages. In K. Gunderson (ed.), *Language, Mind and Knowledge* (3–35). Minneapolis: University of Minnesota Press.

Lipton, P. (1999). All Else Being Equal. *Philosophy*, 74, 155–168.

Locke, J. (1979). *Essays Concerning Human Understanding*. P. H. Niddtich (Ed.) Oxford: Oxford University Press.

Maynard Smith, J. & Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, 246. 15–18.

Michalos, A.C. (1986). Job satisfaction, marital satisfaction, and the quality of life: A review and a preview. In F.M. Andrews (Ed.), *Research on the quality of life* (pp. 57-83). Ann Arbor: University of Michigan, Survey Research Center.

Mill, J. S. (1863). *Utilitarianism*. Chicago, IL: University of Chicago Press.

Myers, D. G. & Diener, E. (1995). Who is Happy, *Psychological Science*, 6. 10–17.

Myerson, R. B. (1991). *Game Theory: Analysis of Conflict*. Cambridge, Massachusetts: Harvard University Press.

Neale, S. (forthcoming). Implicit Reference. In G. Ostertag (Ed.), *The Philosophy of Stephen Schiffer*, Cambridge: MIT Press.

Neale, S. (2007). Heavy Hands, Magic, and Scene-Reading Traps. *European Journal of Analytic Philosophy*, 3, 77–132.

Neale, S. (2005). Pragmatism and Binding. In Zoltan Gendler Szabo (Ed.), *Semantics versus Pragmatics*. Oxford: Oxford University Press, (2005) pp. 165–285.

Neale, S. (1992). Paul Grice and the Philosophy of Language. *Linguistics and Philosophy*, 15, 509–559.

Neale, S. (1990). *Descriptions*. Cambridge: MIT Press.

Nosofsky, R. & Palmeri, T. (1997). An Exemplar-Based Walk Model of Speeded Classification. *Psychological Review*, 104, 266-300.

Pavot, W., Diener, E., & Fujita, F. (1990). Extraversion and happiness. *Personality and Individual Differences*. 11. 1299-1306.

Phillips, J., Misenheimer, L., & Knobe, J. (draft). Love and Happiness. Forthcoming in *Emotion Review*. http://pantheon.yale.edu/%7Ejk762/Love-Happiness.pdf

Popper, K., Kneale, W. C., & Ayer, A. J. (1948). Symposium: What Can Logic Do for Philosophy?. *Proceedings of the Aristotlean Society, supplemental volumes*, 22, 141-178.

Posner, R. (1980). Semantics and Pragmatics of Sentence Connectives in Natural Language. In J. R. Searle, F. Keifer, and M. Bierwisch (Eds.), *Speech Act Theory and Pragmatics*. D. Reidel Publishing: Dordrecht, Holland.

Pratt, J. W., (1964) Risk aversion in the small and in the large, *Econometrica* 32, 122-136.

Putnam, H. (1978). *Meaning and the Moral Sciences*. London: Routledge & Kegan Paul.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, Massachusetts: Harvard University Press.

Récananti, F. (2004). Pragmatics and Semantics. In L. R. Horn and G. Ward (Eds.), *Handbook of Pragmatics*. (442-462). Malden, Massachusetts: Blackwell.

Récananti, F. (1998). Truth-Conditional Pragmatics. In A. Kasher (Ed.), *Pragmatics* (Vol. IV). (512–531). London: Routledge.

Récananti, F. (1998). Primary Pragmatic Processes. In A. Kasher (Ed.), *Pragmatics* (Vol. IV). (512–531). London: Routledge.

Récananti, F. (1986). Defining Communicative Intentions. *Mind and Language*. 1, 213–242.

Rosenthal, D. (2005). *Consciousness*. Oxford: Oxford University Press.

Rosenthal, D. (1997). "A Theory of Consciousness." In Ned Block, Owen Flanagan and Guven Guzeldere (Eds.), *The Nature of Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press, 729-753.

Ross, W. D. (1930). *The Right and the Good*. Oxford: Oxford University Press.

Russell, D. (2009). *Practical Intelligence and the Virtues*. Oxford: Clarendon Press.

Sadock, J. (1978). On Testing for Conversational Implicature. In P. Cole (ed.), *Syntax and Semantics, Vol. 9: Pragmatics* (281–298). New York: Academic Press.

Samuels, W. (1974). You Cannot Derive "Ought" from "Is". *Ethics*, 83, 159-162.

Saul, J. (2001). Review of *Implicature: Intention, Convention, and Principle in the Failure of Gricean Theory* by Wayne Davis. *Noûs*, 35: 630–41.

Saul, J. (2002). Speaker meaning, what is said, and what is implicated. *Noûs*, 36: 228–48.

Schiffer, S. (1986). Compositional Semantics and Language Understanding. In R. E. Grandy & R. Warner (Eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. (175–208). Oxford: Clarendon Press.

Schiffer, S. (1972). *Meaning*. Oxford: Clarendon Press.

Searle. J. (1975) *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Searle. J. (1969). A Taxonomy of Illocutionary Acts. In K. Gunderson (ed.), *Language, Mind and Knowledge, Minnesota Studies in the Philosophy of Science, Vol. VII* (344–369). Minneapolis: University of Minnesota Press.

Searle, J. (1964). How to Derive "Ought" from "Is". *The Philosophical Review*, 73, 43-58.

Seligman, M.E.P. (1991). Learned optimism. New York: Random House.

Silverberg, A. (1996). Psychological Laws and Non-Monotonic Logic. *Erkenntnis*, 44, 199–224.

Singer, P. (1973). The Triviality of the Debate Over "Is-Ought" and the Definition of "Moral". *American Philosophical Quarterly*, 10, 51-56.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Snow, N. (2009). *Virtue as Social Intelligence: An Empirically Grounded Theory*. New York: Routledge.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Sperber, D., & Wilson, D. (1981). Irony and the Use-mention Distinction. In P. Cole (ed.), *Radical Pragmatics* (295–318). New York: Academic Press.

Stalnaker, R. (2006). Saying and Meaning, Cheap Talk and Credibility. In Anton Benz, Gerhard Jäger, & Robert van Rooji (Eds.), *Game Theory and Pragmatics* (83-100). Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Stalnaker, R. (1998). Pragmatics. In A. Kasher (Ed.), *Pragmatics* (Vol. IV). (55–69). London: Routledge.

Stalnaker, R. (1989). On Grady on Grice. *The Journal of Philosophy*, 86. 526-7.

Stock, W.A., Okun, M.A., Haring. M.J., & Witter, R.A. (1983): Age and subjective well-being: A meta-analysis. In R.J. Light (Ed.), *Evaluation studies: Review annual* (Vol. 8, pp. 279-302). Beverly Hills, CA: Sage.

Suppes, P. (1986). The Primacy of Utterer's Meaning. In R. Grandy & R. Warner (eds.), *Philosophical Grounds of Rationality* (109–129). Oxford: Oxford University Press.

Tomasello, M. (2009). *Why We Cooperate*. Cambridge, Massachusetts: MIT Press.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, Massachusetts: MIT Press.

Vranas, P. (2004). John M. Doris, *Lack of Character: Personality and Moral Behavior* [Book Review]. *The Philosophical Review*, 113, 284–288.

Velleman, D. (1985). Practical Reflections. *The Philosophical Review*, 94, 33–61.

Walker, R. (1975). Conversational Implicatures. In S. Blackburn (Ed.), *Meaning, Reference and Necessity* (133–181). Cambridge: Cambridge University Press.

Warneken, F. & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, 311, 1301-1303.

Warner, R. (2001). Introduction: Grice on Rationality and Reasons. In P. Grice, *Aspects of Reason*. (R. Warner, Ed.). (vii–xxxviii). Oxford: Clarendon Press.

Warner, R. (1989). Reply to Barker and Grandy. *The Journal of Philosophy*, 86, 528-9.

Warner, R. (1986). Grice on Happiness. In R. E. Grandy & R. Warner (Eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. (475–493). Oxford: Clarendon Press.

Wilson, D. & Sperber, D. (2004). Relevance Theory. In L. R. Horn and G. Ward (Eds.), *Handbook of Pragmatics*. (607-632). Malden, Massachusetts: Blackwell.

Wilson, D. & Sperber, D. (2002). Truthfulness and Relevance. *Mind*, 111, 583-632.

Wilson, D. & Sperber, D. (2002). Pragmatics, Modularity and Mind-Reading. *Mddn and Language*, 17, 3-23.

Wilson, D. & Sperber, D. (1998). Mutual Knowledge and Relevance Theories of Comprehension. In A. Kasher (Ed.), *Pragmatics* (Vol. IV). (369–382). London: Routledge.

Wilson, D. & Sperber, D. (1986). On Defining Relevance. In R. Grandy & R. Warner (Eds.), *Philosophical Grounds of Rationality* (243–258). Oxford: Clarendon.

Wilson, D. & Sperber, D. (1981). On Grice's Theory of Conversation. In P. Werth (Ed.), *Conversation and Discourse* (155–178). London: Croom Helm.

Yu, P. (1979). On Gricean Program about Meaning. *Linguistics and Philosophy*, 3, 273–288.

Ziff, P. (1967). On H. P. Grice's Account of Meaning. *Analysis*, 28, 1–8.