The Rationalization of Rationality

by
Sheldon Wein
Saint Mary's University

This is a slightly revised version of a paper presented to the 2003 CanadianAssociation for Reductionist Philosophy conference held at the University of Waterloo on October 10th 2003.

Abstract of "The Rationalization of Rationality" by Sheldon Wein

Contractarians prefer to use (a version of) the instrumental conception of rationality. They tend to be suspicious of amendments to this conception—even by those whose credentials as contractarians are unimpeachable. While such suspicion has its virtues, this paper argues that contractarians must accept some amendment to the instrumental conception of rationality.

Papers from the 2003 meeting of the Canadian Association for Reductionist Philosophy
"The Rationalization of Rationality" by Sheldon Wein, Saint Mary's University

Introduction

Contractarians prefer to use a narrow version of the instrumental conception of rationality. They tend to be suspicious of amendments to this conception—even by those whose credentials as contractarians are unimpeachable. While such suspicion has its virtues, I will I argue that contractarians must accept some amendment to the narrow conception of instrumental rationality.

I begin by outlining the minimal conception of a rational agent compatible with moral agency. I show, in standard contractarian terms, why such agents might be attracted to a morality. I then suggest that an appropriate way to select among competing moral theories is to see how well each does in meeting certain widely acknowledged formal conditions for a morality. I outline these conditions and sketch how contractarian theories of morality fare in meeting these conditions. I focus on the sacrifice condition and the compliance problem which it raises. I argue that it is particularly important that contractarians take this condition very seriously (as they have done) because meeting this condition is essential if the theory is to be a candidate as a viable moral theory. I argue that contractarians can only meet this condition in a truly satisfying way if they adopt a non-narrow conception of instrumental rationality.

The instrumental conception of rationality[1]

The history of moral philosophy includes many accounts of the characteristics a person must have in order to be the proper subject of moral rules. Some of these accounts, such as that one must have a will or a soul, are unhelpful for the simple reason that we are less clear about what is involved in having a will or a soul than we are about being a moral agent. What we want, then, is clearly stated minimal conditions one must meet in order to be properly classified as a moral agent. What properties must one have before one can correctly be said to be morally responsible? An individual is a moral agent—that is, someone rational enough to be the proper subject of moral evaluation—if, and only if, she:[2]

- can perceive contingent facts (and distinguish them from necessary truths);
- can make valid inferences and warranted extrapolations from those facts;
- has aims—cares that things are one way rather than another;
- has a positive attitude toward the satisfaction of her aims and a negative attitude toward their frustration;
- can formulate strategies for attaining that at which she aims;
- acts on those strategies to get what she seeks;
- has a consistent preference set.

There may be more to rationality than this, but anyone who does not have these features simply is not rational and, hence, is not morally responsible.[3]

---

[1] This section and the one titled "Formal conditions for a moral theory" are derived from a commentary I wrote on Chris Tucker's papers on contractarianism and constraint presented at the 2003 meeting of the Canadian Section of the International Society for the Philosophy of Law and Social Philosophy held in Halifax, May 2003. They have been revised thanks to comments by Nathan Brett, Dennis Klimchuk, and Chris Tucker.

[2] "What distinguishes a moral agent from a non-moral agent? A minimum condition is that a moral agent's behavior must conform at least closely enough for a probabilistic prediction to a set of strictures whose contents would be recognized in reflective equilibrium as moral, where, furthermore, at least part of the best explanation for the conformity in question appeals to the fact that the strictures are moral. This condition is necessary: if there are no agents that satisfy it, then there are no moral agents." Chantale LaCasse and Don Ross in "Morality's Last Chance" in Danielson 1998, page 340.

[3] I am grateful to Duncan MacIntosh for discussion regarding what should (and should not) go on this list.

The question is whether such agents ought—rationally ought—to adopt a moral code to guide their behavior.

Rational individuals and morality

Lots of reasons have been offered for why rational individuals might want or need a morality—it improves their soul, it is the only (or most likely) path to a better afterlife, it is the only way to be autonomous, it enriches life—but the one that all contractarians prefer is that, when rational individuals interact with one another, their interactions can lead to sub-optimal outcomes. That this is so is best illustrated by the prisoner's dilemma. Rational individuals who find themselves in prisoner's dilemma should, it seems defect. Yet mutual defection is a Pareto inferior outcome. Morality might avoid this outcome, so rational individuals have reason to consider morality.[4] Of course, they should not consider morality in isolation. They will want to consider whether morality is part of the package of devices available to them which best serves rational agents.[5] Does contractarianism, properly understood, include advice to adopt a moral code, the adoption of which will help people avoid the sub-optimal outcomes that their rationality threatens to force upon them?

Evaluating moral theories

Lots of different methods have been offered for evaluating a moral theory. Recently, the most popular has been to follow Rawls's suggestion that moral theories be evaluated in the same way we evaluate (many) other kinds of theories—*viz.*, by seeing if they can be brought into reflective equilibrium with our considered judgment about what theories of the type being evaluated are supposed to do. We cannot helpfully use (an unmodified version of) this methodology here for at least two reasons. First, the content of a contractarian moral code is not yet very well understood.[6] Second, even if we were confident of the content of the moral code contractarianism would yield, it would be difficult to know what to make of the fact, if it were one, that we found it easy to bring the content of the code into reflective equilibrium with our considered moral judgments. This is because contractarians understand their theory to be a radical theory—one which ought to show that many of our considered moral judgments (when considered from a contractarian perspective) are not worthy of our allegiance.[7] Contractarians are surely correct in wanting to allow for the possibility that even some

---

[4] For an argument that altruism will not do the job, see Sheldon Wein, "Prisoners' Dilemmas, Tuism, and Rationality", where it is shown that even perfect altruists face prisoner's dilemmas and hence need either moral rules or an authority solution to help escape the sub-optimal outcome individual rationality seems to recommend.

[5] It is possible that there is a moral code which by itself is the best package for such agents. But for present purposes I will simply assume (along with all contractarians) that whatever the best package is it will include other elements besides morality. (Most particularly, there will be a legal system which functions to offer an authority solution to some prisoner's dilemmas.)

[6] Bob Bright thinks that contractarianism would lead to the same sort of moral code as a preference utilitarian with strong egalitarian leanings would favor, Hobbes thought it led to a totalitarian welfare state, Narveson thinks it leads to the sort of libertarianism favored by the early Nozick, and I think it likely yields the sort of morality favored by bourgeoisie members of a property-owning democracy. I rather suspect that the moral code which each of these thinkers sees as following from contractarian moral theory matches the thinker's antecedent moral inclinations. Coincidence? Or is contractarianism not yet rigorously enough developed to surprise its adherents with the views it generates? Note that contractualism does not face this problem. The differences between, say, Rawls and Scanlon on the public policy output of their theory are miniscule compared with the differences between various contractarians.

[7] When Plato first advances the theory in the *Republic* he seems to advance it as an account of our conception of justice (as descriptive rather than revisionary). But whatever the difficulties in understanding Hobbes's position,

of our most strongly held considered judgments about what morality requires may be unable to withstand rational scrutiny. To put it in Strawsonian terms, contractarians take themselves to be presenting a *revisionary* rather than a *descriptive* moral theory.[8] Consequently, they should not agree to have their thinking evaluated by a standard that favors descriptive over revisionary theories. (Though, as observed in the first footnote in this section, it is somewhat puzzling that when people turn to the task of using contractarian reasoning to generate a moral code the output seems always to be one that the theorist finds attractive *ex ante*!)

Given this I suggest that we evaluate contractarianism not by the *content* of its moral code but by its ability to meet the *formal conditions* any viable moral code must meet. The question, then, is whether contractarianism can be brought into reflective equilibrium with the standardly recognized formal conditions all moral codes are required to meet. This process, which I believe is more likely to be neutral between contractarianism and its competitors than any alternative process available to us, has the advantage of revealing something important about where contractarians need to focus their energies to improve the theory.[9]

Formal conditions for a moral theory

Rational individuals—those who meet the standards set out above and, hence, could be morally responsible individuals—may need a moral code to live by. A moral code is a system of more or less coherent rules which tell one what one ought to do. But not just any set of prescriptive rules counts as a moral code, as a morality. For a system of social rules to count as a morality, it must either meet the following formal conditions or the theorist advocating the set of rules must explain why the condition need not be met.[10]

- the **rationality condition**: the rules of the code must be addressed to rational individuals, for only rational individuals are thought to be subject to the requirements of morality.[11]
- the **ought-implies-can condition**: the rules of morality must be confined to telling the individual to do things that she can do and to refrain from actions from which it is possible for her to abstain.
- the **benefit condition**: the rules of morality must be such that, taken collectively, those who follow them will be better off than they would be if they did not follow them. Any set of social rules which did not make a community governed by those rules better off than it would be without those rules would not be a moral code.[12]

---

it is clear that he saw himself as advancing a subversive account of justice—one that would, if accepted, require a substantial revision of our way of thinking. (For an excellent account of Hobbes *qua* subversive thinker, see Edwin Curley's "Religion and Morality in Hobbes" in Coleman and Morris.) Gauthier and Narveson are also explicit in thinking that contractarianism must be thought of as a revisionary moral theory.

[8] P.F. Strawson, *Individuals* (London: Methuen, 1959). Strawson explains the distinction between descriptive and revisionary *metaphysics* on pages 9-12.

[9] I suspect that this is true of the other main moral theories, but I will not pursue this idea here.

[10] I am grateful to several people at the 2003 CARP meeting for convincing me that these conditions must not be seen as necessary conditions for something being a moral code. For an alternative account of the formal constaints on a moral code, see *A Theory of Justice* §23.

[11] This applies to moral practioners and not moral patients (if the two groups are thought not to be co-extensive).

[12] Of course, not all actual moral codes have been to the benefit of the community at large. (Indeed, perhaps none of them have been!) This fact lends plausibility to the claims against morality made by Thrasymachus, Nietzsche, and Marx. But all moral codes must be such that it is reasonable to suppose that, at some point in time, using the code could reasonably have been seen to be in the interest of some significant segment of the community.

- the **constraint or sacrifice condition**: moral rules are necessarily rules which ask the individual to forgo the pursuit of her own interests or utility (usually for the sake of furthering the interests or utility of others).[13]
- the **categorical condition**: moral imperatives must be categorical, rather than hypothetical. Moral imperatives apply to rational individuals just in virtue of the fact that they are rational individuals. Thus, one is not allowed to beg off the requirements of morality simply because one does not feel inclined to behave in the way morality requires.[14]

the **universality condition**: moral imperatives must apply to all those who meet the rationality condition (as outlined above). Thus a moral code must not arbitrarily exempt some people from the requirement of following its dictates.[15]

Note that moral nihilism is simply the doctrine that there is no logically possible set of social rules which both should win the assent of rational individuals and meetsall these conditions.[16]

A moral theory—one which meets the conditions set forth above—is a theory which succeeds in telling us how to construct a moral code worth having. Its output is a set of rules telling moral agents what they ought to do. It also serves to justify and defend the rules of the code. (Some moral theories do not meet all conditions outlined above. But when they do not, the defenders of the theory usually offer an explanation of why those conditions the theory does not meet cannot or should not be met. And it counts against a moral theory that it does not meet some of the conditions. Thus, a theory which met none of the conditions above would be a non-starter, not even worth our consideration.)

There are substantial advantages to evaluating moral theories by seeing how well they meet the formal conditions outlined above. From my point of view, the most important of these is that it allows for a possibility we need to keep open: that the proper content of a moral code may be at odds with—indeed radically different from—the moral rules we acquired in the course of becoming potential moral agents. Indeed, this approach even allows for the possibility that having a morality is (rationally speaking) a bad idea. Broadly speaking, this last possibility might be for two different sorts of reasons. Perhaps other ways of solving our coordination problems—religion or superstition, say—are better (cheaper to operate and more likely to succeed, say) than morality. Or perhaps solving coordination problems—at least the sorts of coordination problems moralities can solve for us—just is not worth the effort.

A bit of (quick and dirty) history

Contractarians have little trouble meeting the first three conditions on the list above. Indeed, contractarianism does at least as good a job of meeting these conditions as does any alternative moral theory. And, at least since Gauthier, contractarians have been the philosophers who have devoted the most attention to the fourth condition, to stressing the importance of the compliance problem and to

---

[13] By using the expression "interests or utility" here, I am seeking to find a neutral way of putting what is going to be a central issue in this paper.

[14] In an earlier formulation of this condition—see my paper "Plato and the Social Contract"—I confused categoricality with universality. I am indebted to Malcolm Murray both for pointing out this error and explaining to me why it is such an important error to avoid.

[15] Chris Tucker has suggested that this list may not be exhaustive and that both and **impartiality condition** and a **relative stability** condition may need to be added. I have not added these conditions here but my initial thoughts suggest that exploring them will not reveal either any important virtues in, or difficulties with, contractarianism.

[16] Thrasymachus is usually thought to be the first moral nihilist. Nietzsche is the thinker who devoted the most care and energy to making this view a live option for philosophers. For an excellent contemporary defense, see Richard Joyce, *The Myth of Morality*, New York: Cambridge University Press, 2001. J.L. Mackie calls this view "moral skepticism".

trying to solve it. But, of course, they are not the only philosophers to deal with the problem. It may be helpful at this point to sketch a couple of non-contractarian solutions to the problem as an aid to understanding the limits on a contractarian solution.

First, consider Plato's solution to the problem.[17] In the *Republic* Plato offers us the first statement of the contractarian theory of justice. He rejects it and goes on to offer a radically different theory.[18] While Plato scholars dispute how to best characterize the details of his theory, the broad outlines are now well understood. Basically, Plato holds that we need to draw a distinction between our actual or real self-interest and our apparent self-interest. He also holds that any account of human psychology which cannot allow for the possibility that people can come to act on their real rather than their apparent self-interest must be flawed. Hence, he defends a tripartite account of the soul which makes logical space for such action. Contractarians, reject both these moves. They all cleave to a Humean account of action according to which there are but two parts to the soul (if I can use such unfashionable language), beliefs and desires. All action results from a belief/desire pair, and rational action is just that action which maximizes the satisfaction of desires given the beliefs the agent has. Furthermore, contractarians take desires as given, as not being subject to rational evaluation.[19] But note that, once one has taken Plato's route, showing how a moral code can meet both the benefit and the sacrifice conditions is easy. One benefits by being moral because being moral just is in one's real self-interest, and one makes a sacrifice by being moral because being moral is, sometimes at least, not in one's apparent self-interest.[20]

Utilitarians have offered a wide variety of ways of meeting the compliance condition. There is Mill's (in)famous "proof", offered in *Utilitarianism* (and, on some readings at least, the same ideas are to be found in Bentham).[21] In addition, there is R.M. Hare's argument that using moral terminology at all commits one (if one is careful about using the terms as their logic requires) to a form of utilitarianism.[22] And, of course, there is John Harsanyi's derivation of average preference utilitarianism from the equiprobability model.[23] Indeed, Sidgwick's claim in the *Methods of Ethics* that reason cannot tell us how to choose between utilitarianism and egoism can be seen as a sort of answer to the compliance problem. (Utilitarianism is, on this view, tied with egoism in meeting the compliance condition. And, of course, egoism cannot meet the other conditions nearly as well as Sidgwick's utilitarianism can. It is

---

[17] In my view, contractarians have given Plato's thoughts on these matters too little consideration. While I think Plato's own solution to the problem is wildly extravagant, his insights into the nature and depth of the problem are not, I believe, fully appreciated by contractarians. See "Plato and the Social Contract" and "Plato's Moral Psychology".

[18] Plato's rejection of contractarianism seems to be based on its failure to meet the categorical and universality conditions. I take the story of the Lydian Shepherd to be performing this role for Plato.

[19] Hume famously puts this point in forceful terms, pointing out that one can prefer to scratch one's finger rather than not, even when doing so will result in the destruction of the entire world, and that one can prefer one's own lesser good to one's greater good. (Obviously, the latter remark requires some interpretation to avoid it being the contradiction it threatens to be if one's "good" is taken to be just whatever one "prefers".)

[20] Another possibility would be to hold that being moral is the most efficient means to what is in one's actual (real) self-interest rather than it being constituent of one's actual self-interest. Presumably this latter possibility is a bit less intolerable to contractarians.

[21] Bob Bright convinced me of this.

[22] I have phrased Hare's position this way so as to be neutral between his early views and his (much stronger) later views. See particularly his *Moral Thinking* and Allen Gibbard's contribution to Seanor and Fotion.

[23] On this, see Lanning Sowden and Sheldon Wein, "Justice and Rationality: Doubts about the Contractarian and Utilitarian Approaches".

for this reason that Rawls rejects egoism as a viable ethical theory worthy of consideration by his hypothetical contractors.)[24]

Alternatively, consider the Kantian solution. Kant, like the contractarians, thinks that some of our actions are best explained by belief/desire pairs.[25] But he holds that, in addition, we have the capacity to follow rules that we impose upon ourselves. Indeed, like Rousseau, he thinks that full rationality, freedom, and autonomy are only realized when one acts on such rules. Furthermore, he thinks that, as it happens, the one rule one could rationally give oneself is the Categorical Imperative *and* that the Categorical Imperative generates a moral code meeting the conditions specified above. Hence, Kantians take Kant to have solved the compliance problem.[26]

Contractarians and the benefit condition

The official position among contractarians, at least as I understand it, is that none of the above answers to the compliance problem has much to be said for it.[27] That is to say, contractarians hold that other moral theories—whatever the virtues of those theories in meeting the other formal conditions—fail in the task of showing that it is rational to make the sacrifices that complying with a moral code inevitably requires of one.Only contractarianism can succeed in showing both why morality demands such sacrifices of us and why making those sacrifices is something it is rational to do. Put bluntly, contractarians think this because they hold that the prisoner's dilemma models the sorts of problems a morality can solve for us. If this is so, it is instantly clear how a morality can both be of benefit and demand substantial sacrifices.

Hobbes, though he did not have the benefit of contemporary game theory to help him model this problem, was nonetheless the first to see it with any great clarity.[28] He recognized that human interaction is such that, by each of us rationally pursuing our own projects, we can and almost inevitably will reach outcomes all of us will find inferior to ones we could have reached. That is to say, Hobbes realized that the normal circumstances of human beings are such that they will frequently find themselves in prisoner's dilemmas.

One way of understanding Hobbes is to read him as holding that there is no solution to the compliance problem. Thus there is no reason to be moral (which is why we have to hire someone—the

---

[24] *A Theory of Justice*, pages 117-8.

[25] According to Philippa Foot (in "Morality as a System of Hypothetical Imperatives"—she makes this claim in the text just following note # 10), Kant thinks the desire in such cases is always the desire for one's own pleasure. Thus, Kant thinks that (for all action but moral action) humans are psychological hedonists.

[26] Note that Rawls, in assuming that his contractors all have the second of his two moral powers (the sense of justice), simply assumes what Kant laboriously argued for. Scanlon seems to just assume this power also. Thomas Nagel is the contemporary Kantian who has spilled the most ink trying to show that reasons must be more than just instrumental ones.

[27] Figuring out what goes wrong with these various attempts is an interesting exercise. It is, of course, extremely easy to give quick objections to each of these solutions. Plato's conception of rationality, which involves rationally evaluating our ends as well as the means to those ends, is far too extravagant to be taken seriously and is coupled with a moral psychology that strains credibility; Mill's "proof" commits the fallacy of composition; Harsanyi begs the question by adopting the equiprobability model as an undefended starting point; and so forth. But seeking to understand the strengths of these unsuccessful attempts and coming to grips with why they were attractive to such bright philosophers is, I think, something that would amply repay the attention of contemporary contractarians.

[28] Plato seems to have come extremely close to having the insight which Hobbes made so much of. For some speculation on why Plato—who actually offers a rather good sketch of the contractarian theory in the *Republic*—missed what Hobbes was able to grasp see my "Plato and the Social Contract

sovereign—to force us to behave in the way that moral people would behave).[29] And Hume, assuming for the moment that he was a contractarian, solves the compliance problem only by taking for granted that we are (universally?) motivated to take an interest in the interest of others.[30]

Contemporary contractarianism and the move to rationalize rationality

Just as Hobbes's *Leviathan* serves as the Old Testament to anyone interested in contractarianism, so David Gauthier's *Morals by Agreement* is the theory's New Testament. Put in less flowery language, *Morals by Agreement* did for contractarianism what *A Theory of Justice* did for Kantian constructivism or contractualism; it rescued a theory once thought to have little promise and made it into a viable option for anyone seriously considering a moral theory. And, just as since *A Theory of Justice* there are now many contractualists where once there were few, since *Morals by Agreement* there are now some contractarians where once there were very, *very*, few.[31]

Consider the chart below comparing the two theories.[32]

1. Type of theory	Kantian constructivist	Hobbesian contractarian
2. Advocate, leading statement	Rawls, *A Theory of Justice*	Gauthier, *Morals by Agreement*
3. Initial choice situation	original position where parties are behind a veil of ignorance	a state of nature where everyone accepts Gauthier's version of the Lockean Proviso
4. Type of decision problem facing parties	individual choice	game theory (bargaining problem)
5. Rationality of parties (their rational decision rule)	maximin rule	constrained maximization
6. Principles agreed upon	Rawls's two principles of justice	maximin relative benefit/minimax relative concession (MRC)
7. Justification for principles of justice	ensures fairness embodies full Kantian rationality	mutual advantage shows that it can be rational to be moral

Both Rawls and Gauthier adopt the idea that one might generate rules to guide our mutual interaction by asking what rules suitably placed parties would agree to, providing only that all others agreed to such rules also. These rules are the output of the theory—the moral code which the theory

[29] On this reading, Hobbes's Foole has the right (Hobbesian) view because none of Hobbes's responses to his Foole are successful. For arguments that Hobbes had a moral theory, see David Gauthier, "Thomas Hobbes: Moral Theorist" (in his *Moral Dealing*).

[30] On Hume as a contractarian, see David Gauthier's masterful "David Hume: Contractarian" (in his *Moral Dealing*). The strongest non-textual argument against reading Hume as a contractarian (rather than a proto-utilitarian) is, in my view, the fact that Adam Smith (who pretty obviously was a utilitarian) seems not to have heard anything from his good friend Hume about why one should prefer contractarianism to utilitarianism. And Smith's economic research practically set him up to be a contractarian. For (some of) my views on Hume on these matters, see my "Humean Minds and Moral Theory" and "A Humean Theory of Distributive Justice for a New Century". Finally, Hume's sensible knave makes one wonder if he thought he could meet the universality condition.

[31] For an explanation of the distribution of contractarians and contractualists, see Sheldon Wein, "American and Canadian Justice".

[32] Some may disagree with parts of the chart. Jan Narveson seems to read Gauthier as though the adoption of the Lockean Proviso is the principle of justice the contractors adopt. Nothing in this paper hinges on this difference in how to interpret the text of *Morals by Agreement*.

generates. As already indicated, I don't think it is wise (for our present purposes) to delve into the merits of the principles agreed upon by the parties in each theory. Nor, I think, can much be gained (again, for our present limited purposes) by exploring whether the initial choice situation is perspicuous. What we are interested in here is rows 4, 5, and 7 under the Hobbesian contractarian column.

The problem Gauthier's contractors face is a prisoner's dilemma. Each of them recognizes that she will be better off if she and her fellow contractors are able to make and keep an agreement to move from the state of nature into civil society. But each would prefer that all others cooperate in this endeavor and that she gained the resulting benefits without having to pay the costs. Worst of all—even worse than remaining in the state of nature and living life without the benefits of civil society—would be to cooperate in the effort to construct a civil society while others did not keep the agreement to do their share in this endeavor. Thus, for each person, it is the case that she is better off defecting from any proposed agreement no matter what the others do. In the (only sometimes helpful) language favored by game theorists, defecting dominates cooperation. The contractors are in a prisoner's dilemma and it would seem, initially at least, that there is no rational way out.[33]

At this point Gauthier introduces what I think is his most important idea, [34] the idea that the contractors should—rationally should—change their conception of rationality from being straightforward maximizers to being constrained maximizers. Because constrained maximizers manage to cooperate with each other in prisoner's dilemmas, and consequently Gauthier is able to show that it is rational for each person to agree to and abide by the constraints imposed by the contract (those that MRC requires of one) *providing that it is rational to move from being a straightforward maximizer to being a constrained maximizer*.

Thus Gauthier is able to show in a particularly compelling way how the contractarian theory generates moral codes that simultaneously meet both the benefit and sacrifice conditions. The constrained maximizers benefit because their outcome (mutual cooperation) is Pareto superior to the outcome reached by straightforward maximizers. But they still see that there is a sacrifice that each has to make—the rules impose a constraint—for were the others willing to cooperate without such cooperation being contingent on one's own, one could have been a free rider on the efforts to move from the state of nature to civil society, thereby achieving one's best outcome. *Constrained maximizers* are aptly named for it is by adopting a disposition to *constrain* their actions so that they are sometimes able to act in counter-preferential ways (when they are in a prisoner's dilemma with another constrained maximizer) that they are able to choose in a way that *maximizes* the outcomes they can expect from interacting with other equally rational agents. Gauthier's account meets the sacrifice condition by arguing for the necessity for agents who are rational enough to constrain their choices even at the theoretical level. It is for this reason that contemporary contractarianism can be said to better meet the sacrifice condition—to take the constraint problem more seriously—than any other leading moral theory.

Recent debates about contractarianism and rationality

---

[33] If this is so—if there is no rational way to generate rational cooperation based on a set of moral principles, whatever their content—the "contractors" must content themselves with finding a different solution to the problem. Perhaps they could all adopt a set of superstitions or a religion, the belief in which would solve their coordination problem. Or, following Hobbes, they could adopt an "authority solution", an arrangement which changes the payoff structure so that they no longer find themselves in a prisoner's dilemma.

[34] Gauthier seems to agree with me on this matter saying "I remain convinced that this [constrained maximization] is the most fruitful idea in *Morals by Agreement*" David Gauthier "Uniting separate persons" in Gauthier and Sugden, page 185.

Gauthier's development of constrained maximization has—as everyone in CARP knows—spawned considerable interest in the important philosophic task of exploring the appropriate way to conceptualize instrumental rationality. For present purposes I will ignore most of the (very important) subtleties in the literature and develop a rather crude distinction which will, I hope, serve to illuminate an important feature about the ways in which contractarians might best meet the sacrifice condition and show how their theory deals with the compliance problem. I have already gestured towards the distinction I think we need here. On the one hand there are those—Eric Cave, Peter Danielson, Edward McClennen, and Joe Mintoff—who hold that constrained maximization or something broadly like it is the way for contractarians to deal with how to show that it is rational to cooperate in prisoner's dilemmas and, hence, how to show that it is rational to constrain one's preference satisfaction in order to gain the benefits of civil society. On the other side are those—Gilbert Harman, Duncan MacIntosh, and Chris Tucker—who hold that we ought to retain the "traditional" conception of instrumental rationality, according to which (they claim) it is analytic that one must defect in a prisoner's dilemma. Let us call the former group the *rationalizers* and the latter group the *traditionalists*. The rationalizers differ among themselves on how to rationalize rationality. (Very briefly Cave thinks we need to cultivate the capacity to cooperate so that becoming something like a constrained maximizer is part of our natures; Danielson thinks that we to (or have evolved to) be able to forgo our preferences when they conflict with our known interests; Gauthier argues that we should move from being straightforward maximizers in prisoner's dilemmas to being constrained maximizers; McClennen thinks that we should choose resolutely; and Mintoff holds that we should form intentions which resist alteration when no new information is obtained between the moment of intention formation and the moment of action.[35]) But they all hold—agreeing with the traditionalists on this—that unless rationality compels some change in us, instrumentally rational humans are doomed to reach sub-optimal outcomes in their mutual interactions.

The traditionalists all reject the idea that rationality requires us to change our conception of rationality. They hold that we cannot have rational agents choosing counter-preferentially, for such "choice" is, by definition, either irrational or not a choice. So the idea of changing one's conception of rationality and acting according to the new conception (in a constrained manner) is not an option for them. Rather, they all hold that only choice which maximizes preference satisfaction can be counted as rational choice. But, like the rationalizers, the traditionalists differ among themselves on how to get rational agents to choose to "cooperate" in prisoner's dilemmas. (Roughly speaking Harman holds that rational agents need to develop intrinsic preferences in favor of cooperation, or mutual cooperation, MacIntosh holds that such agents must change their existing preference structure so that mutual

[35] I am not sure that I have Cave properly classified here. In some ways his view is similar to Tucker's, so he might be a traditionalist. One might understand Cave as holding that we should do to ourselves what Tucker holds we should do to the next generation. But there is a big difference between re-educating oneself and indoctrinating others. Note that, in broad outline, Gauthier's argument that rational agents need to change their conception of rationality in order to avoid mutual defection in prisoner's dilemmas is like Kant's argument that there is more to rationality than mere instrumental rationality. Of course, there is a huge difference between Gauthier and Kant. Gauthier argues that it is instrumentally rational to become a constrained maximizer while Kant argues that the formal or metaphysical features of rationality or freedom require one to be more than just instrumentally rational if one is to be an agent at all. Danielson's argument is, in a way, more like Plato's. In distinguishing between our interests and our preferences Danielson makes a similar move to Plato when he distinguishes between our real and our actual self-interest. Of course, there are huge differences between Danielson and Plato, the most important being Danielson's strong preference for a minimal metaphysics and Plato's delight in what can only be described as a pronounced programmatic propensity to postulate platoons of Platonic forms.

cooperation becomes the most preferred outcome, and Tucker holds that we need to breed and educate a new generation of conditional cooperators.)[36]

Trashing traditionalists

In the debate between the rationalizers and the traditionalists, I side with the rationalizers. Ideally, I could now produce an argument to show that a particular form of rationalized rationality must be the correct one (and so, consequently, that the traditionalists are wrong). Lacking such an argument, I will instead merely point to the theoretical considerations which, it seems to me, should be sufficient to tip the balance in favor of the rationalizers.

If one follows the traditionalists, the contractarian theory succeeds in meeting the sacrifice condition, but only in a way that does not involve constraint. The traditionalist's rational agents may act in the same way as constrained maximizers do in (what look on not very careful inspection to be) prisoner's dilemmas, but their "cooperation" is not at all constrained.[37] It is what they most prefer to do. So, if the traditionalists hold sway, contractarianism will meet the sacrifice condition only in a very weak way. Given that it already fails to meet my final two conditions, giving into the temptation to skirt rather than meet the compliance problem is damaging to the credibility of contractarians.[38]

Finally, let me just toss out of few other considerations that weigh against the traditionalist position here. First, we should remember that both traditionalists and rationalizers have rejected the economist notion of rational choice in terms of revealed preference. So traditionalists cannot claim that any change to the minimal instrumental conception of rationality is forbidden by general features of the contractarian theory. Both groups correctly realize that if we stuck with the economic conception of instrumental rationality (the choice as revealed preference conception) the contractarian theory could not get off the ground as a normative theory.[39] Second, traditionalists hold that in non-game theoretic contexts rational agents should maximize their expected utility. But, as we know from Mill's paradox of happiness, the best way to maximize one's utility may not be to follow the recommendations of the maximize expected utility principle. Everyone—traditionalist and rationalizer alike—holds that maximizing expected utility is a standard of the rationality of action, not a decision

---

[36] MacIntosh holds that rationality requires rational agents who find themselves in a prisoner's dilemma to change their preferences so that they move from a prisoner's dilemma to an assurance game (though he fails to put it this way). Tucker is more optimistic about education than more "seasoned" educators can allow themselves to be!

[37] "The standard way of thinking about rationality . . . fails to yield a theory that can render intelligible the notion of having a commitment to practice rules, much less provide for the rationality of being so committed." Edward McClennen, "Rationality and Rules" in Danielson, 1998, page 34. This passage is quoted by Chris Tucker in his "Defending the Orthodoxy: Justification and Economic Rationality", 2003.

[38] Here I will just assume that contractarianism generates a system of hypothetical imperatives and that it does not apply universally (as the example of the Lydian Shepherd shows). Chris Tucker has pointed out to me that both these claims require argument. I will only note here that my view is that there are strong reasons for thinking either that we need to understand the categorical condition in a non-Kantian way or that the condition is one that a moral theory need not meet. But I will not explore this matter here. I am unsure of what to say about the universality condition. Finally, let me observe that most non-contractarian philosophers seem to have gotten the impression that contractarians actually take some delight in holding that their theory does not meet these two conditions.

[39] It is because of this rejection that Robert Brandom holds that contractarians like Gauthier (even prior to adopting constrained maximization) have accepted what he labels "minimal kantianism". See Robert Brandom, "What do Expressions of Preference Express?" in Morris and Ripstein.

procedure for acting.[40] So, the traditionalists must either hold that (aside from games like the prisoner's dilemma and problems like the Newcomb Problem and the Toxin Problem) it often is better to be irrational than rational or that rationality does not always require acting in a this-particular-choice-maximizing way. Furthermore, the argument for defecting in a prisoner's dilemma—the dominance argument—has never been shown to follow from the maximize expected utility principle.[41] So, the traditionalists could take some solace in remembering that if they have to become rationalizers and adopt constrained maximization (or something along those lines) as the proper model of choice in prisoner's dilemmas, they are not violating any theorem of rational choice theory.

I conclude that contractarians should seek some form of rationalized rationality and use that form as the basis for enhancing and developing the wonderful work begun in *Morals by Agreement*.[42][43]

[40] No one (except for obstinate undergraduates) holds that utilitarians have to claim that the principle of utility must be used as a decision procedure to be followed at each and every choice point. Rather, we all understand that the principle of utility is a standard of rightness or (as I prefer) a way of ranking possible worlds and that the question of which decision procedure tends to move a community towards more highly ranked possible worlds is an open one. Furthermore, everyone realizes that using *Always seek to maximize utility* is almost certainly among the lousiest ways to move towards more highly ranked possible worlds. Contractarians need not saddle themselves with the same difficultieswith which those who do not understand utilitarianism seek to saddle utilitarians.

[41] In case anyone has concerns here about my metal health, let me say that I am a confirmed two-boxer.

[42] My own view is that the best place to begin this task is with Gauthier's conception of rationality found in his "Rethinking the Toxin Puzzle".

[43] I am grateful to Duncan MacIntosh, Thea E. Smith and Chris Tucker, all of whom discussed these matters with me and made helpful comments (on extremely short notice) on an early draft of this paper. I received numerous helpful comments from several people at the 2003 CARP conference in Waterloo.

Works Cited

Bratman, Michael E.
1987 *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.
1992 "Shared Cooperative Activity" *The Philosophical Review*, Vol. 101. No.2 (April 1992)
Danielson, Peter
1992 *Artificial Morality: Virtuous Robots for Virtual Games*. New York: Routledge.
1998 *Modeling rationality, morality, and evolution*. (editor) New York: Oxford University Press.
Gauthier, David
1969 *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes*. New York: Oxford University Press.
1986 *Morals by Agreement*. New York: Oxford University Press.
1990 *Moral Dealing: Contract, Ethics, and Reason*. Ithaca: Cornell University Press.
1993 "Constituting Democracy" in *The Idea of Democracy*. David Copp, Jean Hampton, and John E. Roemer editors, NY: Cambridge University Press.
Gauthier, David and Robt. Sugden, eds.
1993 *Rationality, Justice, and the Social Contract: Themes from Morals by Agreement*. Ann Arbor: University of Michigan Press.
MacIntosh, Duncan
1991 "Preferences Progress: Rational Self-Alteration and the Rationality of Morality" *Dialogue* Volume XXX No. 1-2, 1991
1992 "Preference-Revision and the Paradoxes of Instrumental Rationality" *Canadian Journal of Philosophy* Volume 22, No. 4, December, 1992
1993 "Persons and the Satisfaction of Preferences: Problems in the Rational Kinematics of Values" *The Journal of Philosophy* XC No. 4, April 1993.
Mintoff, Joe
1993 "Rational Cooperation, Irrational Retaliation" *Pacific Philosophical Quarterly* Volume 74, No. 4, December 1993.
2000 "Is Rational and Voluntary Constraint Possible?" *Dialogue* Volume 39, 2000
Morris, Christopher and Arthur Ripstein
2001 *Practical Rationality and Preference: Essays for David Gauthier*. New York: Cambridge University Press.
Wein, Sheldon
2000 "Global Security and Property Rights"*Humanomics,*Volume 16, # 2, 2000
1998a "A Humean Theory of Distributive Justice for a New Century" *Paideia: Philosophy Educating Humanity* (Papers presented at the Twentieth World Congress of Philosophy and published at conference website)
1998b "Plato's Moral Psychology" *Paideia: Philosophy Educating Humanity* (Papers presented at the Twentieth World Congress of Philosophy and published at conference website) (A Spanish translation of a revised version of this work is also forthcoming)
1997 "Feminist Consciousness and Community Development" *The International Journal of Social Economics*, Volume 24, Issue 12, 1997
1994 "American and Canadian Justice" *Canada: Theoretical Discourse/Discours theoriques*, edited by Terry Goldie, Carmen Lambert, and Rowland Lorimer, (Montreal: Association for Canadian Studies, 1994)
1994 "A Hobbesian Foundation for Welfare Rights" in *The Liberalism-Communitarianism Debate* edited by C.F. Delaney, (Lanham, MD: Rowman & Littlefield, 1994)
1988-89 "Humean Minds and Moral Theory" *Philosophy Research Archives*, Volume XIV, 1988-89

Papers from the 2003 meeting of the Canadian Association for Reductionist Philosophy
"The Rationalization of Rationality" by Sheldon Wein, Saint Mary's University

1986 "Moral Skepticism and Moral Truth" *Philosophy in Context*, Volume 16, 1986
1986-87 "Plato and the Social Contract" *Philosophy Research Archives*, Volume XII, 1986-87
1986 "Problems with Contractarianism" *Journal of Social Philosophy*, Volume XVI, #3, 1986