TOXIC DISPOSITIONS IN STRATEGIC RATIONAL CHOICE

by

KENNETH RAY CROSSLEY

B.A., Trinity Western University, 1991

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

Department of Philosophy

We accept this thesis as conforming
to the required standard

The University of British Columbia

April 1996

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Philosophy

The University of British Columbia
Vancouver, Canada

Date April 16, 1996

Abstract

David Gauthier argues that in order to be rational, agents must accept voluntary constraints on strategic behaviour. These constraints define an agent's strategic disposition. Taking Gregory Kavka's toxin puzzle as a foil, Section One demonstrates how strategic dispositions face two challenges posed by standard accounts of rational choice: (1) since they potentially rationalize particular acts which are not immediately utility-maximizing at the time those acts are undertaken, 'rationally irrational' internal constraints are incoherent; and (2) a rational agent might not be able to adopt the required constraints.

Against the first objection, Section Two exploits the contention of standard rational choice theory that the rationality of actions is best evaluated instrumentally. Natural mechanisms of agency are therefore relevant filters on an agent's rationally-feasible options. Moreover, rational agency can be well interpreted with a naturalistic model of intentional action. Intentional agents are capable of planning. A particular act undertaken to further a broader plan will then be a rational act if the plan is a utility-maximizing plan. A structure of rational plans thus informs a coherent account of strategic dispositions.

Section Three notes that agents could still be unable to adopt Gauthier's internal constraints if they entertained conflicting intertemporal preferences. However when overall-utility maximization is demonstrably more rational than discrete-utility maximization, internal conflicts can be resolved. The requisite priority of overall-utility maximization is established with a pragmatic conception of normative justification. Accounts of rational choice, given their basis in primitive fact, therefore ought to endorse Gauthier's internal constraints on strategic choice.

Table of Contents

List of Figures

Introduction. Moral Philosophy as Social Science

In *Morals By Agreement*,[1] David Gauthier argued that in order to be rational, agents must adhere to self-imposed constraints on their social behaviour. This paper defends Gauthier's claim from objections that strategic dispositions are conceptually incoherent and that rational agents will not be able to adopt the required constraints. In section two, the argument turns on a naturalistic account of rational actions. The argument in section three further invokes an evolutionary conception of justification to guide interpretations of rational choice. This brief introduction suggests that the pragmatic intuitions which drive these narrow arguments are well grounded in important methodological considerations from broader social and political theory.

In the early part of the twentieth century, moral philosophers were preoccupied with clarifying the status of moral propositions and the meanings of moral terms. Since they analyzed linguistic behaviour instead of simply prescribing social norms, their endeavours in 'metaethics' were in a sense descriptive. These early descriptivists worried that since there are no moral facts built into the 'fabric of the universe,' moral talk and moral debate are at least dangerously subjective or relativistic if not altogether meaningless. But none of the resulting emotivism, prescriptivism, or related non-cognitivisms yielded fruitful research programs. Moral theory stagnated.

In the early 1970s, sparked largely by John Rawls' *A Theory of Justice*,[2] hopes for productive ethics revived. Rather than striving for a renewed metaethical cognitivism, Rawls was largely indifferent to the putatively dangerous issues lurking behind moral language. Clarifying his theory of justice as 'political not metaphysical', he instead proposed that

> what justifies a conception of justice is not its being true to an order antecedent and given

> to us, but its congruence with our deeper understanding of ourselves and our aspirations, and our realization that given our history and the traditions embedded in our public life, it is the most reasonable doctrine for us.[3]

That is, for example, *A Theory of Justice* presumed that those of us steeped in liberal democratic traditions tend to value fairness and liberty as components of justice and ways of life worth defending. Whether or not liberty and fairness reflect transcendant, timeless, universal truths is irrelevant to our ways of doing politics. A philosophy which justifies our particular social order with some deep metaphysical *telos* is likewise irrelevant to political argument. Credible moral and political theory will instead include fairness and liberty as standards of justification in social and political philosophy because they play such roles in the moral environment at hand. Rawls' simple insight, then, was to build his moral and political theory on a reflective description of moral and political practice.[4]

Of course social values, principles, and practices can be described in numerous ways. And having undermined access to an 'order antecedent and given to us,' there is tremendous space in which to debate whether some descriptions are better than others. Nevertheless, the absence of an absolute or independent standard with which to arbitrate between competing descriptions need not spawn viciously unbridled relativism. Proposing, testing, and refining alternative descriptions is the standard fare of the social sciences. And a growing number of social scientists recognize that denying recourse to an absolute authority is problematic only if the goal of inquiry is to describe an absolute truth. When the goal is to offer substantively helpful insights, social and political theory still thrives as an effort in artifice.

In much this light, Gauthier characterizes his own moral philosophy as "an attempt to

allay the fear, or suspicion, or hope, that without a foundation in objective value or objective reason, in sympathy or sociality, the moral enterprise must fail."[5] Like social theory in general, moral theory becomes a morals by artifice. The artifice must of course be constructed with resources on which we can rely; but the artifice must be constructed only with resources on which *we* can rely.

In this regard, Gauthier noted that rational choice theorists of social behaviour are beginning to yield concise explanations and testable predictions with increasingly elegant results. Concerned to produce substantive conclusions, these social scientists spurn the 'true' for the useful, the 'real' for the salient. They deal in primitive facts -- beliefs, desires, preferences, outcomes -- with a powerful calculus of rational choice.

Refining the trend which began with Rawls, these tools help rejuvenate moral theory. The proof of this claim is in the results. Gauthier's analysis of 'rational compliance' problems grounds his contention that (in some cases) morality is a necessary condition of rationality. Section One introduces these basic issues in rational choice theory and charts Gauthier's case for rational constraints.

Sections Two and Three defend Gauthier against two possible objections from rational choice theorists. The naturalistic approach to rational action in section two and the pragmatic conception of justification in section three both capitalize on the basis of social science in primitive facts. That is, both arguments presume that theories of rational choice, like the moral theories they inform, should strive for helpful formal descriptions of a world much like our own. Using rational choice theory to derive substantively moral conclusions, we can bolster the credibility of moral philosophy as social science.

Section 1. Toxic Dispositions in Rational Choice

In discussing everyday events we frequently say that actions are 'rational' if it somehow 'makes sense' for people to do those actions. Pioneered for use in descriptive economics, formal theories of rational choice interpret, refine, and schematize intuitions regarding how that which makes sense does make sense. These tools are proving to be increasingly useful for analyzing problems in social and political philosophy. This section explores one such important case -- namely, the rationality of individual compliance with endeavours in collective action.

Subsection one outlines a general apparatus of strategic rational choice and frames the compliance problem. Subsection two endorses David Gauthier's strategic dispositions as internal constraints which make compliance rational. But subsection three, introducing Gregory Kavka's toxin puzzle as an expository foil, poses two potentially significant tensions between Gauthier's strategic dispositions and standard accounts of rational choice: that a rational agent might be unable to adopt a strategic disposition, and that the required constraint is normatively incoherent. These problems set the agenda for sections two and three.

1.1 To expedite discussion, we first sketch basic tools. [6]

Rational choice theorists standardly assume a subjective account of value. A descriptive position, value-subjectivism holds that there is no absolute or universal teleology or good that agents ought to desire or pursue. Consequently, when we wonder what it makes sense for an agent to do, the answer largely depends on what it is that she wants. Given her interests, we can say what actions make sense for her to do to achieve those interests. With a subjective account of value, rationality thus becomes instrumental. The rational actions for any given agent are the

actions which promote the agent's interests.

In order to evaluate the rationality of different alternatives, we analyze interests and preferences with an abstract metric of *utility*. The early roots of utility theory are found in the following St. Petersburg paradox.[7] Suppose a fair coin will be repeatedly flipped until it comes up heads. The game will end with that n^{th} toss, and you will be rewarded a $\$2^n$ prize for playing. You must decide how much it is rational to pay to play the game. Essentially, you face a lottery which affords $\$2^n$ with probability 2^{-n} for each n; that is, there is a ½ probability of winning \$2, a 1/4 probability of winning \$4, a 1/8 probability of winning \$8, and so on. Since the expected payoff of the lottery is infinite (the expected payoff of the lottery is (½)2 + (1/4)4 + (1/8)8 +...= 1+1+1+...), it seems as though you should be willing to pay an unlimited amount to play the game. But you also realize, for example, that you have only a 1/128 chance to win even \$128; paying an unlimited fee could therefore seem silly. Hence the purported paradox: it seems both rational and irrational to pay an unlimited amount to play the St. Petersburg lottery.

The eighteenth century mathematician Daniel Bernoulli invoked the 'moral worth' of money to assess the value of the St. Petersburg lottery. 'Moral worth' encapsulated the plausible intuitions that \$1 is worth more to a pauper than to a wealthy person, and that the first \$100 is worth more to a person than is the next \$100. In the St. Petersburg game, the relative worth of absolute gains decreases as wealth increases while the relative worth of absolute losses increases as wealth decreases. Even though the absolute payoff increases as n nears infinity, the moral worth of additional gains decreases while the moral worth of losses increases. Based on the moral worth of the lottery, Bernoulli argued, it will be rational to stop well before paying an unlimited sum to play the game. More formally stated, the wagers and payoffs in the St.

Petersburg paradox exhibit the property of diminishing marginal utility. 'Moral worth' and diminishing marginality contain intuitive cornerstones for modern rational choice theorists' conceptions of utility.

Though the sources of utilities vary from agent to agent, rational agents can nevertheless sort their prospects into individual preference orderings. For example, while both may prefer oranges to all other fruits, Eric may derive greater utility from apples than from pears while Bruce may derive greater utility from pears than from apples. Eric's preference ordering would then be oranges>apples>pears; Bruce's preference ordering would be oranges>pears>apples.[8] More generally, preference orderings rank options by utility, where some option x affords greater utility for an agent than does option y if and only if she prefers x to y. Where x or y is a lottery, the value of each lottery is the expected utility. The expected utility is the sum of the value of each possible outcome multiplied by the probability that that outcome will obtain.[9] More intuitively stated, if an agent faces two gambles, the lottery with the greater expected utility is the gamble that the agent would rather take.

In a seminal work of rational choice theory, John von Neumann and Oscar Morgenstern demonstrated that given certain technical requirements (which need not detain us here), it is possible to formally represent an agent's utility function.[10] The utility function captures both the ordering of an agent's preferences and the relative weight that the agent assigns to each of the various prospects. That a precise utility scale is arbitrary is no more problematic than it is problematic that either Fahrenheit or Centigrade scales can be used to measure temperature.

Given an agent's utility function, rational choice theory becomes especially interesting when agents cannot be certain of the outcome(s) of any given choice. The rationality of an

action taken under uncertainty is then evaluated as a lottery across possible outcomes. The rationality of a choice among actions is evaluated as a lottery across possible actions. In a lottery across actions, an agent is said to choose a strategy.[11] An individual strategy is a lottery over a single agent's choices. A joint strategy is a lottery over the products of strategies of more than one agent (or a lottery over possible outcomes).

When the only independent variables in determining an outcome are the agent's choices, we say a decision situation is parametric. In parametric conditions, an agent chooses rationally if and only if she acts to maximize her (expected) utility.

When the outcomes available to an agent depend in part upon the choices made by some other agent(s), we say a decision situation is strategic (or interdependent). There are three commonly held conditions on rational strategic action: (1) each agent's choice must be a rational response to the choices she expects others to make, (2) each must expect every other agent's choice to satisfy condition 1, and (3) each must believe her choice and expectations to be reflected in the expectations of every other agent. Notice, however, that each agent's choice across possible actions is a strategy. And notice further that strategies are chosen on the basis of the utility derived from the expected outcome(s) of the possible actions. It is the outcomes, not the strategies, that afford utilities. The difficulty then emerges that in satisfying condition 1, it is not always clear whether an agent trying to maximize utility ought to respond to another's strategy or to another's utility.[12]

This distinction is most important when we consider conditions of equilibrium and optimality. A strategic outcome is in equilibrium if and only if it is the product of strategies each one of which is a utility-maximizing response by each agent to the strategy or strategies

chosen by the other agent(s). We can determine that an outcome is in equilibrium if no agent could increase his utility by unilateral departure from his chosen strategy. But an outcome is optimal if and only if it is the product of strategies each of which is a utility-maximizing response by each agent to the utility or utilities achieved by the other agent(s). We can determine that an outcome of interdependent action is optimal if there is no other possible outcome that yields at least one agent greater utility and no agent less. In strategic conditions, identification of an agent's rational action with his utility-maximizing action is not so clearly cut as in parametric conditions since in any given strategic situation the equilibrium outcome(s) might be mutually exclusive with the optimal outcome(s).

To clarify the apparatus and to illustrate this difficulty which strategic instances can pose for rational choice, consider the following strategic decision problems (or games). Each agent (or player), Row and Column, chooses to act in accord with either strategy A or strategy B. The utilities afforded by each outcome (the payoffs) are for Row and Column respectively.

Figure 1.1a Easy Coordination

		Column A	Column B
Row	A	2,2	0,1
	B	1,0	0,0

Figure 1.1b Chicken

		Column A	Column B
Row	A	3,3	2,4
	B	4,2	1,1

In Figure 1.1a, there is an equilibrium (2,2) where each player's best response to the other's strategy is also an optimal outcome. Strategy A is each player's best response to the other's best strategy. The easy rational solution to the strategic decision problem is the joint strategy AA.

In Figure 1.1b, a variation of the game called Chicken, there are two optimal equilibria, (2,4) and (4,2), and a third optimum, (3,3), which is not an equilibrium. Strategy B is Row's best response to Column's strategy A but not to Column's strategy B, while strategy B is Column's best response to Row's strategy A but not to Row's strategy B. Hence there are individually rational strategies but no rational joint strategy.

Now consider a third case, the famous Prisoner's Dilemma (PD), in which the divergence of the optimum from equilibria generates rational compliance problems.[13] Row and Column independently choose, without an enforceable contract, whether to cooperate with the other or to defect (D).

Figure 1.2 Prisoners' Dilemma

		Column C	Column D
Row	C	R,R	S,T
	D	T,S	P,P

T>R>P>S

If the other cooperates, each stands to gain either R from mutual cooperation or else T from defection. Since T>R, if the other cooperates it pays to defect. If the other defects, each stands to gain either P for defection or else the 'sucker's payoff,' S. Since P>S, if the other defects it pays to defect. No matter what the other does it pays each agent to defect. But each does worse if both choose their individually best response to the other's strategy than each could do if both choose the optimal joint strategy.

We can generalize the logic of interaction which produces the dilemma in this case. It

is individually irrational to participate in mutually beneficial collective action. The optimal joint strategy (CC) affords each greater utility than is possible from both defecting; but each agent's choosing C is not a utility-maximizing response to the other's choice of strategy C. In any situation where it is individually non-utility-maximizing to comply with a joint strategy affording optimal utility, we face a compliance problem.[14]

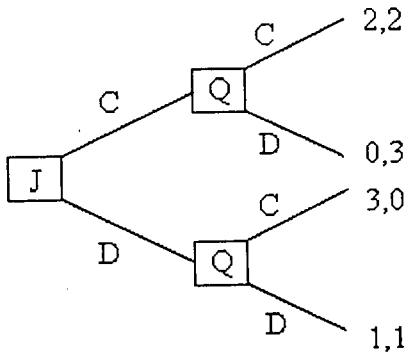
1.2 One suggested solution to our compliance problem is for players to somehow alter their preferences such that they prefer mutual cooperation.[15] Row and Column might each come to value neighbourliness, for example, such that defecting against the other's cooperation is less preferred than is cooperating against the other's cooperation. Adjusted for some value of neighbourliness, R is preferred to T. However we must reject this move as a solution to the compliance problem on the grounds that suitably altering preferences simply changes the structure of the problem to an easy coordination problem rather than solving the dilemma.[16]

Alternatively, players might seek constraints on their unaltered preferences such that compliance becomes rational. In *Morals By Agreement*, David Gauthier proposes just such constraints. Constraints can be external or internal.[17] Like a civil society instituting Hobbes' sovereign Leviathan, like nations enforcing treaties, like citizens organizing property laws, rational agents might construct external devices to restrain or punish non-compliance.[18] However external constraints will likely often prove costly and fail to maximize utility.

We can show that Gauthier's internal constraints solve the compliance problem more efficiently with the following farmers' game.[19] This two-stage sequential decision problem, like the normal form prisoner's dilemma, leads to the suboptimal (1,1) equilibrium under standard

conditions of strategic choice.

Figure 1.3 The Farmers' Game



If Farmer J cooperates, Farmer Q rationally defects. Recognizing this eventuality, J must defect to earn 1 rather than 0; Q must therefore defect rather than himself getting suckered.

But in the sequential game the nature of Gauthier's solution to the compliance dilemma is intuitively clearest. Let Q be precommitted to cooperate if and only if J cooperates. If J defects, Q does too and both get 1. If J cooperates, Q does too and both get 2. Wanting 2 more than 1, J will cooperate if she knows that Q will constrain his interests in 3. In accepting constraint, Q complies with the joint strategy rather than pursuing his individually best strategy given J's cooperation. More importantly, Q complies rationally because without his accepting constraint he will get 1 rather than 2.

Farmer Q's precommitment to conditional cooperation instantiates a 'strategic disposition.' Q is neither changing his preferences nor simply disguising his interests in order to induce J's cooperation. Since he still prefers 3 over 2 but constrains his pursuit of 3 in order to achieve 2 rather than 1, Q's disposition is a robust constraint. Because Q's disposition enables him to achieve a greater reward than would have been possible for him to achieve without his

precommitment to constraint, Gauthier terms a disposition like Q's 'constrained maximization.'

More carefully generalized, a constrained maximizer (CM) conditionally seeks to maximize her utility given the utilities of those with whom she strategically interacts. A straightforward maximizer (SM) is disposed to maximize her utility given the strategies of those with whom she strategically interacts. Thus, in the terms of 1.1 above, constrained maximization makes optimality a necessary condition of strategic rational choice. In strategic conditions, an agent acts rationally if optimizing expected utility.

There are two conditions on rationally constrained maximization: Q rationally cooperates only if J does, and Q rationally cooperates if J does. That is, a constrained maximizer is disposed to cooperate with others who are disposed to cooperate but still to defect against straightforward maximizers; and a constrained maximizer actually does cooperate with other cooperators.

To see how constrained maximization is utility-maximizing, consider an agent participating in several distinct one-shot prisoner's dilemma interactions.[20] If I am disposed to straightforward maximization, I might expect utility u" from choosing my best individual strategy against CMs, and u from choosing my best individual strategy against other SMs. In a population with ratio p CMs to SMs, the SM disposition affords the expected utility pu"+(1-p)u. If I am disposed to constrained maximization, I can expect utility u' from cooperating with other CMs, and u from defecting against SMs. In a population of ratio p CMs to SMs, the CM disposition affords the expected utility pu'+(1-p)u. Since u">u' for any population with some CMs, the SM disposition apparently affords greater expected utility than does the CM disposition.

Of course there is a problem in this construction. It is assumed that when faced with

CMs I successfully achieve u''>u'. But by the first condition on rational constraint, a CM will not cooperate with an SM. As an SM I therefore cannot expect u''>u' when dealing with CMs; I can at best expect u''=u.

So if I am an SM detected by CMs, I expect u from defection against other SMs and u from defection against CMs. SM affords pu+(1-p)u. If I am a CM, I expect u' for cooperating with other CMs and u for defecting against SMs. CM affords pu'+(1-p)u. Since u'>u for all populations with some CMs, the CM disposition for conditional cooperation affords greater expected utility than does SM.

We find this argument persuasive and endorse Gauthier's conclusion (at least for the strategic society of only CMs and SMs engaged in multiple one-shot interactions) that those disposed to constrained maximization do better than do those disposed to straightforward maximization. In differently constituted populations, different dispositions might of course do better. Given these findings, we more generally conclude that a strategic disposition constituting internal constraint is rational. In appropriately qualified strategic conditions, internal dispositions generate rational compliance.[21]

1.3 Developing an account of dispositions broaches conceptual tensions between strategic dispositions and orthodox accounts of rational choice. This subsection raises two of the latter troubles -- namely, problems of conceptual incoherence and problems of illicit substantivism.

Standard decision theory evaluates choices with respect to particular decisions maximizing expected utility. In contrast, Gauthier's dispositions are rational metastrategies; it is not particular strategies or actions that merit utility-maximizing justification, but the

overriding strategy by which subsequent strategies and actions are chosen. Moreover, these metastrategies restrict the set of actions an agent is even able to perform. For example, consider the claim that the best disposition a rational agent could have is to follow 'honesty as the best policy' but to take advantage of exceptional circumstances. Gauthier counters that a constrained maximizer "is not able, given her disposition, to take advantage of the 'exceptions.'"[22] In like manner, a rationally constrained Farmer Q is not able to defect having induced J's cooperation with a precommitment.

Reversing this notion that an agent's dispositions curtail his abilities to perform certain acts, it makes sense to wonder whether a rational agent could be unable to adopt a strategic disposition of constraint. The objection emerges clearly from Gregory Kavka's toxin puzzle. An eccentric billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink a vial of toxin tomorrow. The toxin will make you very ill for a day but will have no mortal or otherwise lasting effects. You need not actually drink the toxin to get the million; you need only intend to drink the toxin. The billionaire will not mistake your intention. External contraptions to prevent your not-drinking (thus persuading the billionaire of your sincerity) are not options. Neither is it possible for you to decide 'to intend tonight but change your mind in the morning,' for then your intention to drink is clearly not your intention at all. But after having received your prize, it seems irrational for you to drink the toxin. You face the toxin puzzle: "You are asked to form a simple intention to perform an act that is well within your power...You are provided with an overwhelming incentive for doing so. Yet you cannot do so."[23]

The toxin puzzle captures a potential difficulty with Gauthier's internal solution to the

compliance problem as represented by the farmers' game. Farmer Q's sacrifice of 3 for 2 by cooperating even after J cooperates parallels your drinking the toxin even after having received the prize. Farmer Q's precommitment to constraint is analogous to your sincere intention tonight to drink the toxin tomorrow. You and Q would both do well to have appropriate dispositions guiding particular subsequent behaviours; but you and Q both know that the relevant dispositions would require acts which are not utility maximizing at the time they are performed. Just as you can not commit to drinking the toxin, Q could be unable to adopt a disposition of constraint.

To generalize our analysis, let us say that choices regarding utilities afforded by possible outcomes at a given decision instance are discrete choices which afford discrete utilities. Aggregated discrete-utilities obtained by an agent from discrete choices constitute that agent's overall-utility. The rationality of a disposition derives from considerations of overall-utility. However the toxin puzzle makes it clear that governing individual decisions with devices of overall-utility can require instances of behaviour that are discretely-irrational. Apparently beneficial metastrategies become deceptively toxic dispositions.

This central tension between discrete and overall rationality frames two significant challenges for a defence of strategic dispositions. Firstly, inabilities to adopt a disposition might stem from incompatible conditions on rational choice. If they render rational particular acts which are clearly not utility maximizing when those acts are undertaken, or if they render irrational acts which clearly are utility maximizing when those acts are undertaken, then strategic dispositions do seem as though they might be conceptually incoherent. This is the objection to strategic dispositions from 'rational irrationality.'[24]

Notice, however, that the truly puzzling aspect of the toxin puzzle concerns the

relationship of an agent's rational beliefs and intentions to an agent's rational actions. That is, the toxin puzzle and the charge of rational irrationality potentially depend on a particular conception of rational agency. The rational inability to adopt a disposition might then be a function of the intentional structure of a rational agent. This recognition invites a conceptual tie between a theory of rational agency and the coherence of rational irrationality -- namely, that an agent capable of intrapersonally coordinating rational choices will be able to adopt strategic dispositions. We will develop this claim in Section Two.

The second challenge posed for dispositions by tensions between overall and discrete utilities concerns the degree to which the argument for dispositions relies on misplaced substantivism. Substantiating the normative rationality of strategic dispositions requires somehow prioritizing possible 'long-term' benefits over 'short-term' losses. But granting value-subjectivity, it could seem as though an agent might simply prefer maximizing discrete-utilities to maximizing overall-utility. Prescribing overall-utility maximization against the temptations of discrete-utility maximization potentially smuggles an illicit value-objectivism into rational choice. Moreover, even if the charge of objectivism can be answered, it could still seem that the 'overall' and 'discrete' frameworks require more context than is appropriate given the rigorous abstraction which usually drives theories of rational choice. In Section Three we develop and rebut these challenges with an account of the pragmatic justification of normative conditions on rational choice.

Conclusion:

David Gauthier's strategic dispositions solve the rational compliance problem at the risks

of an incoherently rational irrationality and of illicit substantivism. For Gauthier's strategic dispositions to enhance a viable theory of rational choice, rational agents must be capable of adopting intrapersonal constraints. The supposed problem of rational irrationality plays on a potential inability of rational agents to adopt dispositions of internal constraint. An inability to adopt a disposition could stem either from the objective constitution of the agent or from antecedent rationality constraints. Section Two will argue that in order to sustain the toxin puzzle, Kavka employs a faulty conception of rational agency. Repairing the model of agency defends Gauthier's dispositions against the objective inability aspect of rational irrationality. Section Three will then defend a principle of overall-utility maximization as an antecedent constraint on rational choice. Overall-utility maximization relies on a pragmatic conception of normative justification to preserve the coherence of dispositions without relying on illicit substantivism.

Section 2 Active Agents, Idle Intentions

This section complements rational choice theory with considerations of intentional action, isolating aspects of rational agency which invoke natural causal mechanisms as filters on an agent's set of feasible actions. The resulting model enables us to reject Kavka's interpretation of the toxin puzzle and leads to the conclusion that if it is rational to intend to drink the toxin then it will also be rational to drink the toxin.

Subsection 1 generalizes the frameworks of deliberation Gauthier and Kavka respectively need to plausibly support their conclusions in the toxin puzzle. Subsection 2 shows how these competing accounts derive from a deeper disagreement regarding the relationship of rational intentions to rational actions. Subsection 3 proposes a causal-intention model of rational agency as a tool with which to resolve the disagreement. Taking Alfred Mele's assessment of the toxin puzzle as a foil, subsection 4 examines the relevance of natural causal mechanisms to the rationality of intentions and actions. Subsection 5 begins generalizing these findings with an intuitive account of rational plans. Subsection 6 tightens these intuitions with decision-theoretic apparatus. Subsection 7 reformulates the toxin puzzle with the refined tools and restates the stakes of rational deliberation in the toxin puzzle such that Kavka might still escape rationally intending to drink the toxin even when rational intentions are linked with rational actions.

2.1 Faced with Kavka's toxin puzzle, Gauthier argues that not only is it rational for an agent to intend to drink the toxin, it is rational for her to actually drink it. Kavka argues that a rational agent will not be able to intend to drink the toxin. Kavka's and Gauthier's dispute stems from different characterizations of the agent's alternatives in the toxin puzzle. In order for his

conclusion to carry *prima facie* plausibility, Gauthier needs the agent to deliberate between drinking the toxin or else not-receiving the prize. If the agent can win the prize without drinking, that is clearly his best option. But if the agent evaluates the disutility of not-winning against the comparatively painless drinking, winning the prize and drinking is the utility-maximizing option. For these to be the stakes of deliberation, since it is the intending and not the drinking that actually earns the reward, the drinking of the toxin must come packaged with the intention to drink.

Alternatively, in order for his reading to carry *prima facie* plausibility, Kavka needs the agent to deliberate between not-intending to drink the toxin or else drinking the toxin. If the agent measures the disutility of drinking the toxin against the comparatively painless status quo of not-intending to drink, not-intending to drink is clearly the utility-maximizing option. However if the agent measures the utility of not-drinking the toxin against the disutility of losing out on the million, not-drinking the toxin is clearly not the utility-maximizing option. Kavka's account of deliberation will be wrong if drinking the toxin and winning the million come in a package. These two will come packaged as long as rationally intending to drink the toxin somehow commits the agent to drinking the toxin. To substantiate the inability of a rational agent to intend to drink the toxin, Kavka must show that Gauthier's packaging of the stakes of debate is mistaken.

2.2 While neither Kavka nor Gauthier accepts that the rationality of performing some act implies the rationality of the intention to perform that act, Gauthier holds the converse while Kavka severs all ties. These general differences can be articulated in the narrow terms of the toxin

puzzle.

For referential convenience, let the agent facing the toxin puzzle be named Cindy. That Cindy's drinking the toxin tomorrow would be irrational does not imply that Cindy's intending to drink the toxin tomorrow would be likewise irrational. Assuming that she wants her life to go as well as possible, and assuming that the million dollars will help further this aim, it is rational for Cindy to intend to drink the toxin even if it would not be rational for her to act upon that intention once her prize is effectively won. More generally, performing an irrational act need not imply that the intention to perform that act is irrational. On this point, Kavka suggests, he and Gauthier concur.[25]

Kavka worries that Gauthier endorses a converse implication from the rationality of intentions to the rationality of actions. By Gauthier's own account, *A* is an intentionally restrictive act if and only if an agent choosing *A* becomes faced with a choice among possible acts some of which are intentionally incompatible with her *A*-ing.[26] That is to say, if some set of actions, *S*, which would be available to an agent at a later time (t_2) if she does not choose *A* at an earlier time (t_1) would be intentionally incompatible with choosing *A* at t_1, then at t_1 act *A* is intentionally restrictive of act(s) *S* at t_2. If an agent identifies an intentionally restrictive act, she must decide which act(s) would make her life go best among those intentionally compatible with *A*. When deliberating upon her course of action, an agent therefore cannot rationally commit herself to *A* if she finds that *A*-ing will intentionally restrict her from *B*-ing while she intends to at least include *B* among her options. For her to keep *B* among her options, she must think it possible for her to perform *B*, and must intend to perform *B* should her deliberations inform her that *B* is her best option. Moreover, if she intends today to *A* tomorrow, then other

things being equal, she cannot rationally choose tomorrow not to do that which she intends today.[27]

Since it will advance her presumed aim that her life go better, Cindy rationally forms the midnight intention to drink the toxin. By Gauthier's account, the rational action in the morning is the action that will make her life go as well as possible subject to the constraint that it be intentionally compatible with her previous commitment. Drinking the toxin is intentionally compatible with the night's deliberation while not drinking the toxin is not similarly compatible; drinking the toxin is therefore rational.

By Kavka's account, Cindy has very good reason to intend to drink the toxin. But when the time comes, Cindy will nevertheless have very good reason not to drink the toxin. Because the rationality of the intention is insufficient for generating the rationality of the corresponding action, it will be irrational for Cindy to drink the toxin.

In advancing this account, Kavka denies that the rationality of an intention implies the rationality of its corresponding action. Instead he maintains that intending to perform *A* could be rational because of desirable effects from the intending, while the actual *A*-ing could still be irrational because of undesirable effects from the acting. Moreover, if an agent determines today that to *A* tomorrow will be irrational tomorrow, then an agent cannot rationally intend today to *A* tomorrow. For Kavka, an intention can thus acquire an ambiguous status, both rational as a utility-maximizing act in its own right and irrational as the intent to perform a non-utility-maximizing act. Reasons for intending and reasons for intentionally acting diverge. When this divergence obtains, Kavka finds conflicting standards of rational evaluation: one set for rational actions, another for rational intentions, and still another for evaluating 'the agent's own

rationality.'[28]

Kavka acknowledges something unsettling about this 'ambiguous' even 'paradoxical' conclusion, but prefers the unsettled air to endorsing the evident irrationality of drinking the toxin after obtaining the prize. Gauthier retorts that it is "mad not to be the sort of person who would drink the toxin," and that rationality gives coherent expression to the aims of those who are able to intend to drink.[29]

Neither Kavka nor Gauthier explicitly recognizes that their underlying dispute concerns antecedent rationality commitments brought to a shared structure. We intimate this conclusion by noting that both positions can be captured with the same framework of deliberation. Gauthier's central claim need essentially be only that where later reasoning is subordinated to prior intent, otherwise irrational acts become rational. Where deliberations at some time t_1 are intentionally incompatible with actions at some later t_2, then since deliberations at$_1$ t take normative priority, the deliberative framework restricts that which an agent may rationally do at t_2. Working within the same framework, Kavka's alternative claim need be only that where prior reasoning is subordinated to later action, otherwise rational dispositions become irrational. Where deliberations at t_1 are intentionally incompatible with actions at t_2, then since actions at t_2 take normative priority, the deliberative framework restricts that which the agent may rationally intend at t_1. Deciding which, if either, restriction is rational requires further explanation of the 'normative priority' of temporally distant deliberations. That task will occupy the greater part of our efforts in section three.

But to reach the ground of normativity, we must first show (against Kavka) that there is no fragmentation in standards for evaluating intentions and actions. For at this point it does

seem that Kavka could catch Gauthier on a simple error in logic. Let P be 'it is rational to intend to perform act *A*'; let Q be 'it is rational to perform act *A*.' Gauthier and Kavka both deny that Q implies P. Furthermore, the fact that performing *A* is irrational need not imply that the intention to perform *A* is irrational; not-Q does not imply not-P. However Gauthier then seems to argue that since it is rational to intend to drink the toxin, it is further rational to drink the toxin. In other words, Gauthier apparently argues that P implies Q. But if P implies Q, Kavka could straightforwardly object, it clearly follows that not-Q implies not-P. Our reading of Gauthier's argument seems to have committed him to an internal inconsistency.[30]

To address this objection, it is important to note that ' implications' between rational intentions and rational actions are not fully captured in the sentential logic. Speaking carefully, Gauthier and Kavka agree that performing an act which does not in and of itself contribute to an agent's utility need not imply that the intention to perform that act is irrational if the intention to perform that act does contribute to an agent's utility. However, and it is this point that the above objection fails to take seriously, 'having the intention to perform the act' and 'performing the act' are not so easily separated as the language of the debate so far seems to have suggested. When Gauthier appears to be asserting that P implies Q, the Q in question is somewhat different from the Q in the denial of 'not-Q implies not-P.' In the denial, Q is best read as 'it is rational to perform the act of drinking the toxin in the morning.' In the assertion that P implies Q, Q is better understood as 'it is rational to continue the action begun by forming the intention before midnight to drink the toxin in the morning.' That is, the Q in the assertion 'P implies Q' is not so much an action in its own right as it is the continuation of an act initiated earlier. Kavka might worry that it does not make sense to talk about the rationality of such a non-agentive Q.

But rather than supporting an objection against Gauthier, this potential worry of Kavka's is exactly the point; Gauthier does not need to talk about the rationality of the non-agentive Q. Rather, Gauthier only needs to show that a rational agent should drink the toxin in the morning. And if rationally intending to drink the toxin and rationally drinking the toxin constitute an objectively unified action in a way that the distinct actions of rationally intending to drink the toxin and rationally not-drinking the toxin cannot, then it is a confused objection to require separate justifications for the intention to drink and the subsequent drinking.

Gauthier's condition of 'intentional compatibility' provides the necessary unifying device. In the next two subsections we see that 'intentional restrictions' can arise by virtue of the causal mechanisms which define a structure of rational deliberation. That is, by examining the role of a causal-intention model of rational agency, we will see that intending to drink the toxin (intending to *A*) is an intentionally restrictive act such that not-drinking (not-A-ing) is no longer an objectively possible option available to a rational (intentional) agent.

2.3 By isolating metaphysical aspects of agency, we are not overstepping the bounds of this debate. Both Kavka and Gauthier implicitly acknowledge that the constitution of the intentional agent is important in rationality evaluations. Their respective dependencies on a metaphysics of agency are interestingly evident in their reactions to David Lewis' treatment of deterrence.[31] Citing examples from deterrence theory, Lewis sides with Kavka against Gauthier in noting that 'intending to *A*' is a distinct action from 'intentionally *A*-ing.' Nevertheless, in departure from Kavka, Lewis does not see that the irrationality of *A*-ing therefore restricts an agent from rationally intending to *A*. Lewis instead maintains that a policy containing certain acts might be

rational even while carrying out those very acts could be irrational. Kavka objects to Lewis that the parameters of debate are set by intentions so robustly construed as to exclude the policies of nations from counting as relevant intentions.[32] Similarly, Gauthier worries that Lewis' account of rational agents fails "to express the unified concern that characterizes an individual, that distinguishes him from a mere aggregation."[33] Both Kavka and Gauthier deny membership in the class of rational agents to Lewis' nations on the grounds that agency requires a rather more human psyche than Lewis invokes. We will suggest below that such exclusivity is unwarranted since the very model which plausibly connects rational choice theory with an account of intentional agency warrants a more inclusive conception of agency. But for our broader purposes, it is especially important only that both Kavka and Gauthier do relate rational choice to the metaphysics of intentional agents.

The nature of this relation can be made broadly coherent with a 'causal-intention' model of intentional agency in the tradition pioneered largely by Donald Davidson.[34] The causal-intention model features the central premise that an act is an intentional act or is done intentionally if and only if there is a causal relation between an agent's desires and beliefs such that the reasons for an agent's *A*-ing are the agent's desire to achieve goal G and her belief that she might obtain G by *A*-ing. In the causal-intention model, an agent's intentions are 'conduct controllers.' An agent does not control her intentional actions separately from her intentions; rather, an agent's control of her actions goes by way of her intentions.

This model of intentional agency is amenable to important demands on a model for rational agency. An act is rational or is done rationally if and only if there is an appropriate relation between an agent's preferences and beliefs such that the reasons for an agent's rationally

A-ing stem from the agent's desire to maximize her utility and the agent's belief that she might do so by *A*-ing.[35] An agent does not separately control her rational intentions and her rational actions; rather, an agent's control of her rational actions goes by way of her rational intentions.

A causal-intention model of rational agency extends the causal-intention model of intentional agency in assigning a normative element to causal mechanisms between belief-desires and actions. This normative assignment subordinates the rational to that which is metaphysically possible. This subordination is not especially startling considering that rational choice is instrumental. That which it is rational for an agent to do is a subset of that which the agent is able to do.

The causal-intention model of rational agency underwrites each position on the toxin puzzle.[36] Kavka's account presumes that an agent's belief that-not-*A* (i.e. that it is not rational to drink the toxin) preempts a causal relation from her beliefs to her desires appropriate for manufacturing the intention to *A*. Gauthier's account presumes that an agent's desire for *A*, or for rewards unobtainable without *A*-ing, provokes a revision in her beliefs that-not-*A* and enables her intention to-*A*.

We need not endorse either of Kavka's or Gauthier's speculations regarding what sort of agents can and cannot have rational intentions in order to exploit the causal-intention model of rational agency. Indeed, we reject metaphysical realism with respect to folk-psychological entities such as intentions or beliefs or desires, and instead accept intentional functionalism. Functionalism ascribes belief-desire agency to whatever systems are sufficiently complex to exhaust a mechanistic-causal model for explaining observed events.[37]

Endorsing a functionalist interpretation of the causal-intention model of rational agency

has harsher implications for Kavka's position than for Gauthier's. Gauthier's insistence on the agent as a unity rather than an aggregation is in keeping with the spirit of taking an intentional stance and with the related causal-intention model of rational agency. In order to reasonably ascribe agency to some system, the components of that system need to be sufficiently complex and sufficiently integrated such that the functioning of that system is best explained in terms of the behaviour of an integrated whole rather than in terms of the behaviour of each component part. Where such complexity exists it is not unreasonable to speak of that system as a unity rather than as an aggregation.[38]

However Kavka's objection to Lewis in effect requires that we interpret intentions in terms of human folk psychology. This move exceeds the metaphysics required to adopt the intentional stance. If we can adopt the intentional stance toward an agent, then we can ascribe some modicum of rational behaviour to that agent. That agent need not have a human psyche. The causal-intention model of rational agency only depends upon our being able to define and predict a causal relation between functionally defined beliefs and preferences and behaviour.

With this functionalist model of rational agency we are now in position to defend derivations of rational actions from the rationality of controlling intentions. By invoking considerations of the causal-intention model of rational agency, we will in effect use Kavka's own tool against him.

2.4 We take Alfred Mele as a foil for developing our argument that natural causal mechanisms have bearing on the rationality of intentions and actions.[39] Mele argued that a uniquely disabled agent, Ted, can win Kavka's million. Mele defends Ted's success on the basis

of 'ability-sensitive reasons' to act. The abilities in question concern the agent's control over intentional actions.

Imagine that there is an evil genius who causes Ted to drink toxins whenever he comes across them unless Ted will drink the toxins on his own. Ted is aware of his affliction. Approached by Kavka's billionaire, Ted knows that either he will intentionally drink the toxin or else, by the devise of the evil genius, he will unintentionally drink the toxin. Mele argues that since Ted knows he has no reason not to drink the toxin, he lacks reason not to sincerely intend to drink the toxin. Ted wins the million and intentionally drinks the toxin.

We cannot (nor do we want to) deny Mele's conclusion that it is rational for Ted to intentionally drink the toxin. We do balk at Mele's particular use of ability-sensitivity to establish this result. A defender of Kavka could successfully object that Ted's ability-sensitivity is irrelevant to his rational reasons for acting. Even though Ted knows he will end up drinking the toxin he still has reason not to drink the toxin because it will make him ill. And since Ted still has reason not to intend to drink the toxin even though he will drink it against all his best intentions, he will still be unable to rationally intend to drink the toxin and will rationally end up drinking unintentionally.

Mele counters (i) that even though it is not open to Ted whether he does or doesn't drink the toxin, it is open to him whether he drinks intentionally or unintentionally. And (ii) given these options, there is 'no point' in Ted's intending not to drink the toxin since there is a payoff for intending to drink and because he knows he will drink whether he intends to or not. From these two considerations we are to infer that Ted will rationally intentionally drink the toxin.

Kavka preemptively challenged a version of (I) when, in framing the toxin puzzle, he

noted that intentions are not internal commands which an agent can simply summon at will. Intentions are rather a function of the agent's reasons to act. Kavka's point clearly and plausibly presumes the causal-intention model of rational agency. Furthermore, it sets the onus on Mele to demonstrate that Ted really may 'pick' drinking intentionally rather than unintentionally.[40]

Mele's case turns on the notion that an agent can form an intention sufficient for intentional *A*-ing without having a reason to *A*. His (reconstructed) argument runs as follows: (P1) In 'normal' agents, possessing the beliefs that there is no reason to *A* and that there is a reason not to *A* renders the agent incapable of forming an intention to *A*. (P2) Due to his 'abnormal' inability to not-drink, neither intentionally nor unintentionally drinking is either intrinsically nor instrumentally valuable to Ted. (P3) From P2 we infer that even given P1, Ted might choose whether or not to drink on grounds other than the utility afforded him from drinking itself. (P4) The prize is "an excellent reason for 'picking' intentional toxin drinking which is not also a reason to drink the toxin (nor to drink it intentionally)."[41] Ted might then use the prize as the extrinsic grounds with which to pick intentionally over unintentionally drinking the toxin without having a reason to drink the toxin.

But Kavka's advocate may object that the million dollars cannot be extrinsic grounds for Ted's rational decision. Grounds for rational action, as Mele's own argument suggests, are 'extrinsic' if they are in no wise intrinsically or instrumentally valued but if they determine whether or not to do an act or to produce ends about which the agent is indifferent. Consider a modern ennui-ridden philosopher indifferent regarding whether to be or not to be. Disabled by mortality, she knows that whether she wants it to or not, her life will end. It is nevertheless open to her whether to end her life intentionally or else to wait for her life to end unintentionally. In

this case, she might decide by the toss of a coin whether or not to end her life. Since a coin toss could not influence whether or not to value intentionally ending her life, a coin toss is an extrinsic ground for deciding whether or not to intentionally or unintentionally die. But now let our philosopher discover some desire for ending her life intentionally rather than unintentionally. That a coin toss served grounds for choice in the first case does not come to bear in the second case. The second philosopher has grounds to intend to die such that a coin toss becomes inappropriate.

Ted's grounds for decision parallel those of the second philosopher. Because intending yields the prize, Ted attaches utility to intentionally drinking the toxin even though he knows he will end up unintentionally drinking the toxin should he fail to muster the relevant intention. Thus (contrary to P2) Ted instrumentally prefers intentionally over unintentionally drinking the toxin. And since intentionally drinking is instrumentally valued by virtue of the prize it obtains, then (contrary to P3) the prize cannot be an extrinsic rather than 'intrinsic' ground for picking between intentionally or unintentionally drinking the toxin. And since the prize is an intrinsic ground, then it is no longer open even to the uniquely disabled Ted to choose to intentionally rather than unintentionally drink the toxin. That is, there is no choice at all for Ted to make: the conjunction of his belief (that the intention carries instrumental worth) and his desire (to receive the million dollars) precludes his believing that it is open to him whether to intentionally rather than unintentionally drink the toxin. Mele's crucial consideration (I) misses the mark.

Fortunately for our case against Kavka, the failure of (I) can be explained with reference to causal mechanisms which must inform a theory of action. The toxin puzzle is Mele's foil to target aspects of the causal-intention model of agency. But Mele's treatment of 'normal' agents

in P2 simply is the causal-intention model of agency. If we use the model instead of targeting it, we can bolster Mele's failed notion of an agent's intentionally A-ing without having reasons to A. In so doing, we establish the rationality of Ted's drinking without relying on any bizarre causal abnormality to make the case.

Ted's intentional structure is 'abnormal' in that, by the devise of the evil genius, he believes that he will do an act which he ordinarily would not intend to do. But though the circumstances of his unintentional drinking are indeed bizarre, this intentional structure is not especially abnormal. I could quite normally believe, for example, that when I walk in fresh snow I will make tracks even if I have no intention whatsoever of making tracks. If I am being pursued by nasty villains through snowy terrain, I believe that in fleeing I will, due to normal causal mechanisms, do something (i.e. make tracks) which under the circumstances I would prefer not to do.

Now suppose I am being pursued by rogues through fresh snow, and suppose that I must run since there is no place to hide. I then have reason to run in fresh snow without having reason to make tracks (and indeed having reason to not-make tracks). But if I take any time to reflect on the predictable consequences of my action, I will know that I cannot intentionally run through snow without believing that I will make tracks. Because running causes tracks, and because I intend to run while knowing that running causes tracks, it is not a stretch to say that by intending to run I am (regretfully) choosing to intentionally make tracks even if I do not intend to make tracks. More generally, since *A* causes *B*, and because an agent intends to *A* while knowing that *A* causes *B*, it is not a stretch to say that the agent is (regretfully) choosing to intentionally *B* without intending to *B*.

By the same token, it would be rational for Ted to intentionally drink the toxin even without there being an evil genius. Whether it is rational for me to intentionally make tracks depends on whether or not my running is more rational than my not-running. If it maximizes my utility to not-make tracks rather than to run, then I have reason not to intentionally make tracks (and not to run). But if it maximizes my utility to run rather than to intentionally not-make tracks, then I do have reason to intentionally make tracks (without intending to make tracks) because intentionally running causes tracks. The evil genius causing Ted to drink is analogous to tracks being caused without running. This circumstance might indeed influence me to run in the first place (unless the tracks lead away from me) and thus whether it is rational for me to intentionally make tracks. But our point is that this circumstance does not bear independently upon the intending to run and upon the making of tracks but only on the two together.

Because he overlooked the role of natural causal mechanisms in defining an agent's intentional structure, Mele thought he required an external device to coordinate beliefs and desires sufficiently to drive his case in the toxin puzzle. This thought is plausible if the grounds for Ted's rational intention are severed from the grounds for Ted's rational action. Thus we find Kavka's distinction between grounds for evaluating rational actions and grounds for evaluating rational intentions lurking behind Mele's use of the evil genius.

However we showed that with the causal-intention model, an intermediary is not required to drive Mele's case. In so doing, we indirectly deployed Kavka's causal-intention model to argue against the generalizability of his own distinction. Our rebuttal has been, in effect, that there are cases where since *A* causes *B* and it is rational to *A* then one rationally intentionally *B*'s

(even if it is not rational to intend to *B*). If intending to drink causes drinking and it is rational to intend to drink then one rationally intentionally drinks.

This discussion begins to make clear, in a way that the stand-off between Gauthier and Kavka in 2.2 above did not, that the rationality of an action does not follow simply from an abstractly generalized logic of rational intentions but in part from the causal-intentional structures which define rational behaviour.

2.5 Stated in terms of rational choice theory, subsection 2.4 showed that natural mechanisms of intentional action pose filters on the set of options available to a rational agent. We begin incorporating these limited results into our broader argument with Michael Bratman's account of rational planning.[42] Bratman charts a two-level structure of practical reasoning where (1) prior intentions and plans pose deliberative problems and filter solutions to those problems; and where (2) 'desire-belief' reasons enter practical reasoning as the considerations weighed when deliberating between options for action. This two-level structure of practical reasoning draws pragmatic justification from its long-term contribution to planning agents getting what they want. We will deal with the idea of pragmatic justification in section 3. For now we are interested in the structure of planning.

Bratman argues that plans must be both internally consistent and consistent with the agent's beliefs. Should either of these conditions fail, planning ceases to contribute to the agent's thriving in her environment and thereby violates the pragmatic justification condition of practical reason. Plans must also satisfy a condition of means-end coherence. If a plan is conceived but the means of instantiating the plan cannot be filled in, then that plan becomes means-end

incoherent. Moreover, plans are different from desires. Desires are permissibly inconsistent. For example, I could desire both to play basketball and to type my thesis today, even while knowing that I cannot do both. In contrast, desires do not require satisfaction of means-end coherence. For example, I could desire to play basketball without settling on means to satisfy that desire, but if I plan to play basketball then I do need to fill in sufficient means to that end.

Since plans must be coherent, the having of a plan poses further deliberation problems. The agent must first decide between alternative or even conflicting means of fulfilling the plan. And since plans must be consistent, the having of a plan constrains the having of further plans in so far as the means for fulfilling alternative plans cannot conflict with the earlier plan. If two plans are found to be inconsistent, an agent might reconsider one or both of them in light of the desires she wishes to fulfill. Plans therefore constitute a framework for determining which options are relevant and admissible to an agent without providing reasons in favour of one admissible alternative over another. The prioritizing reasons come from the planning agent's desires. Plans and intentions, by structuring the process of weighing 'desire-belief' reasons to act, provide 'framework reasons' for an agent to act.

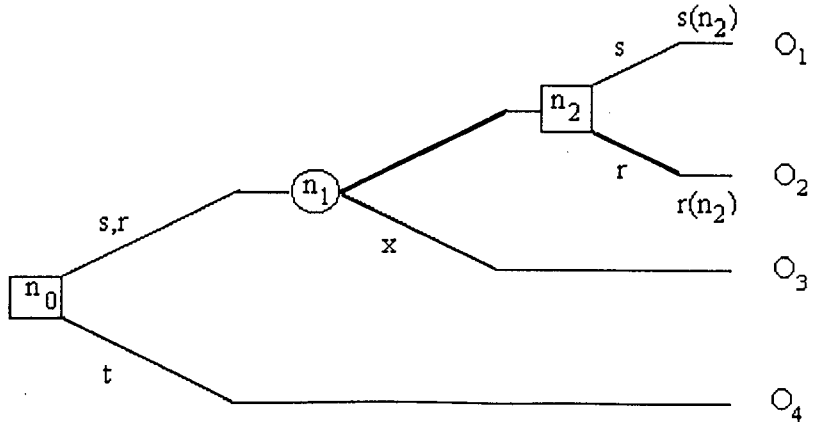
In the broad terms so far stated, we find little in Bratman's two-level structure of rational plans with which to quarrel. We do note that Bratman is at the same time remarkably accommodating and unduly limiting. He suggests that plans may be partial. For example if I plan to play basketball tonight, I do not need to immediately decide how to get to the court and whose team I shall try to be on. And he suggests that plans exhibit a hierarchical structure. This is to say only that plans about ends might embed plans concerning means, and that one could deliberate about parts of plans while holding other parts of the plan fixed. Taken together,

'partial' and 'hierarchical' plans range so widely that they might include wishes, ambitions, dreams, boasts, and all manner of similar future-regarding desires.[43]

Yet despite this permissiveness, Bratman also suggests that a plan is "a certain kind of mental state, not merely an abstract structure of a sort that can be represented, say, by some game-theoretical notation".[44] This limiting characterization of plans is problematic. The consistency and coherence requirements that Bratman's 'framework reasons' provide for rational actions are powerfully developed in decision theory -- an area of analysis built in part around abstracted 'game-theoretical notation.' The two-level structure of deliberation is enhanced when articulated with the very tools of abstracted decision theory that Bratman eschews. For this purpose, in the next subsection we consider Edward McClennen's synthesis of standard decision-theoretic notation.

2.6 In Section 1.2 we represented the farmer's game with a simple sequential diagram schematizing Q's decision structure. The general dynamics of deliberation regarding plans are well represented with similar decision theoretic apparatus.[45] Let T be any bounded sequential decision problem such as the one in Figure 2.1 below.

Figure 2.1 Dynamic Consistency in Rational Plans[46]



Each decision point is a node. An agent's decisions are represented with squares, and decisions taken by nature or chance with circles. An outcome (or terminal node) completes a defined set of choices and chance events. A defined path of choices and events from some node to an outcome is a plan.

Bratman plausibly suggested that some plans might be 'partial.' Let us define S as the set of plans available to an agent facing T, and *s*, *r*, *t*, and so on as the members of S. Let us further define $T(n_i)$ as a truncated tree from the i^{th} node, n_i, to whatever outcome(s) might be reached from that node. A partial plan is that part of a broader plan continuing along a truncated tree. For example $s(n_2)$ is a partial plan in $T(n_2)$ of the broader plan *s* in T.

Bratman suggested that plans might be 'hierarchical.' We note that when reaching any given node, some plans are no longer options for the agent while other plans may continue from that node. For example choosing either *s* or *r* from n_0 moves the agent through the decision tree toward n_1, thereby precluding plan *t*. Also, any given plan may follow the truncated continuation of several distinct plans in S. For example plan *s* follows the truncated continuation of both plans $s(n_1)$ and $r(n_1)$.

Bratman suggested that plans be internally consistent. Let us further define D(S) as the subset of plans in S that are deemed acceptable by the agent, $D(S(n_i))$ as the plans available at n_i from the vantage point of n_i, and $D(S)(n_i)$ as those plans continuing from n_i. The distinction between the two latter sets is worth explanatory emphasis. $D(S)(n_i)$ is the set of acceptable plans available at n_i viewed from n_o; that is, the set of plans deemed acceptable before any moves are made in the tree either by the agent or by chance. On the other hand $D(S(n_i))$ is the set of plans deemed acceptable by the agent from the vantage point of n_i; that is, the set of plans still deemed

acceptable after moves have been made through the tree. We can now define a sense of internal consistency. Following McClennen's articulation, we say a plan is dynamically consistent if and only if for any choice point n_i in T, if $D(S)(n_i)$ is not empty and $s(n_i)$ is in $D(S(n_i))$, then $s(n_i)$ is in $D(S)(n_i)$; and if $s(n_i)$ is in $D(S)(n_i)$, then $s(n_i)$ is in $D(S(n_i))$.[47] This is to say that if an agent adopts some plan at a given choice point and the plan is possible at that node, then the plan must continue from that node; and if the plan an agent adopts continues from a choice point, the plan must be possible at that node. Dynamic consistency simply requires that agents adopt a plan to follow from a particular node that is possible to follow from that node.

We might illustrate with a case of dynamic inconsistency. Suppose in the above Figure 2.1 that an agent evaluating the options available to her from n_0 judges both *t* and *r* to be unacceptable and chooses *s*. Plan *s* can obtain only with a move toward n_1 and, in the chance event that she arrives at n_2, choosing $s(n_2)$ rather than $r(n_2)$. She moves toward n_1 planning to pick $s(n_2)$ if the opportunity arises. Now assume she gets to n_1 and regards $r(n_2)$ as the only acceptable choice despite her earlier intentions to choose $s(n_2)$. There is an evident inconsistency between what she intended to choose at n_2 from the perspective of n_0 and what she ended up choosing upon arriving at n_2. We have a case where $s(n_1)$ is in $D(S)(n_1)$ but is not in $D(S(n_1))$ since $D(S(n_i))=\{r(n_i)\}$. The agent's deliberation fails the test of dynamic consistency.

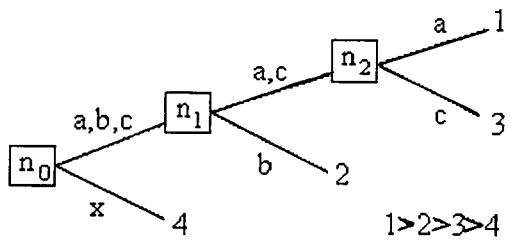
Now recall that rational deliberation requires evaluating the possibility of various options. Since rational choice is instrumental, if it is not possible for some agent to *A*, it is *a fortiori* not rational for that agent to *A*. The subordination of the rational to the possible is captured with notions of feasibility. Feasibility in turn captures and enhances Bratman's notion of means-end coherence and internal consistency. Moreover, again following McClennen, our decision-

theoretic apparatus further clarifies a fine-grained distinction between objective-feasibility and rational-feasibility.

A plan is 'means-end coherent' if it is objectively-feasible (o-feasible). A plan is o-feasible for an agent at some choice node if (a) the agent can reach that node and (b) it is something she can do at that node, given all the natural and technological constraints on her situation.

An o-feasible plan is further rationally-feasible (r-feasible) if it does not require that the agent make a choice contrary to the rationality constraints to which he is committed. R-feasibility includes consistency with the agent's beliefs. Let an agent be committed to a rationality constraint, TR. TR requires that if he prefers a to b and b to c, then he must prefer a to c. Now consider figure 2.2 below, where *a*, *b*, *c*, and *x* are plans leading respectively to the outcomes most preferred to least preferred. At n_0 the agent opts against *x*, enabling a subsequent choice at n_1 between *b* and either of *a* or *c*. Opting against *b* enables a further choice at n_2 between *a* or *c*.

Figure 2.2 R-Feasibility

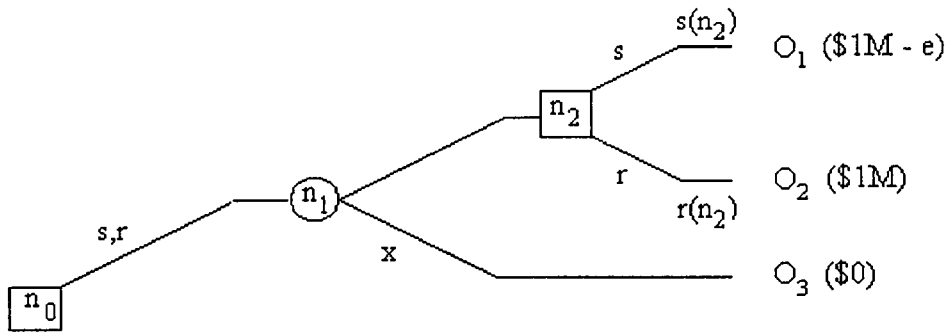


Although objectively feasible, plan *c* is not r-feasible from the vantage of n_0 once given the agent's commitment to TR and the belief-desire to opt against *b* at n_1.

Given this decision-theoretic apparatus for rationally feasible plans, we are equipped for another look at the toxin puzzle.

2.7 Kavka maintains that it would be rational to intend at midnight to drink the toxin since the action of intending would have the desirable consequence of winning the million. However he also maintains that it would be irrational to drink the toxin in the morning since that action would have only ill consequences. Gauthier objects that treating deliberations in this discrete manner presents agents with the self-defeating task of rationally outwitting their own rationality.[48] For Gauthier, rational deliberations are not discrete but are embedded within a broader framework of intentional compatibility. This broader framework has been well represented with our apparatus for dynamically consistent feasible plans.

Figure 2.3 The Decision Structure of the Toxin Puzzle.



Let $s(n_2)$ and $r(n_2)$ in figure 2.3 be to drink and to not-drink the toxin once the prize has been awarded. The node n_1 represents the billionaire's assessment of Cindy's intention.[49] Let the disutility of drinking the toxin be converted to monetary terms by some factor *e*. The outcome O_1 afforded by plan *s* is then \$1M - *e*. The outcome Q afforded by plan *r* is \$1M. Failing *r* or *s*, the agent neither gains nor loses when reaching O_3.

Noting the objective conditions stipulated in order that n_1 affords objectively feasible continuations toward n_2 (i.e. that the agent adopt plan $s(n_0)$), it is clear at n_0 that $D(S)(n_2)=\{s(n_2)\}$. However if Cindy actually reaches n_2, $D(S(n_2))$ will then be $\{r(n_2)\}$. Gauthier's requirement that Cindy's deliberations in the morning be intentionally compatible with her deliberations in the evening is then the requirement that her choice at n_2 be part of a consistent feasible plan chosen at n_0. In opting for $s(n_2)$ at n_2 rather than for the $r(n_2)$ which she actually prefers at n_2, Cindy consistently follows the plan s that is clearly utility maximizing given the terms of continuation afforded at n_1 when viewed from n_0.

Gauthier's claim that the rationality of Cindy's intention to drink the toxin implies the rationality of her drinking can be expressed with the conditions (a) that not-drinking the toxin at n_2 (i.e. choosing plan r) is dynamically inconsistent with choosing plan s at n_0, and (b) that it is rational at n_0 to choose s rather than r. Taken together, (a) and (b) suggest that given the rationality constraints to which Cindy is committed, r(n2) is not r-feasible. The intending to drink and the drinking, rather than being two distinct actions, constitute one consistent plan. Satisfying a rationality constraint that the agent maximize expected utility requires optimizing among the options O_1, O_2, and O_3. But O_2 is not a feasible plan. Therefore winning the prize (improving on O_3), comes only by packaging intending to drink (choosing $s(n_2)$) with intentionally drinking (choosing $s(n_2)$ rather than $r(n_2)$).

A defender of Kavka still might object that just as Gauthier's requirement of intentional compatibility and his logic of intentions can be recast in terms of dynamic consistency among feasible plans, so can Kavka's counter-position to Gauthier. Because she deems $s(n_2)$ unacceptable (i.e. because she believes it is irrational to drink the toxin once the prize is

awarded), and because she cannot plan inconsistently to fix on $s(n_0)$ but choose $r(n_2)$, Kavka's Cindy finds herself forced to consistently opt for O_3 rather than either *r* or *s*.

Kavka's claim that the rationality of Cindy's drinking the toxin is not implied by the rationality of her intending to drink the toxin can then be expressed with the conditions (a) that choosing plan $s(n_2)$ is dynamically inconsistent with endorsing plan $r(n_2)$ at n_0, and (b) that it is rational at n_2 to choose *r* rather than *s*. The conjunction of these premises is tantamount to the assertion that given the rationality constraints to which Cindy is committed, $s(n_2)$ is not r-feasible.

From these results, Kavka's advocate could suggest that the toxin puzzle retains its force to the extent that Gauthier's intentional compatibility requirement and Kavka's 'divergent standards' rebuttal are both r-feasible. That is to say, we might conclude that dynamically consistent deliberation among feasible options is an incomplete logic of deliberation for resolving the toxin puzzle.

We will throttle this objection at its source in section 3. For now it is enough to note that our decision theoretic model of the toxin puzzle assumes that O_1 rather than O_2 is the terminal node of a sincere intent to drink the toxin. The objective constraints posed by the causal-intention model of rational agency guarantees this result. Because the intention comes packaged with the action, we can use short hand for the intending to drink and the drinking.[50] The options in the toxin puzzle become 'winning the prize having drunk,' 'winning the prize having not-drunk,' and the appropriate negations of each. These events apparently yield three interpretations of an agent's alternatives in the toxin puzzle:[51]

(1) Cindy deliberates between receiving the prize having drunk the toxin or not receiving

the prize having not-drunk the toxin.

(2) Cindy deliberates between receiving the prize having drunk the toxin or receiving the prize without having drunk the toxin.

(3) Cindy deliberates between receiving the prize without having drunk the toxin or not-receiving the prize having not-drunk toxin.

Of course the plurality of even just these three is illusory. By the thought experiment which Kavka laid down, the only viable puzzle for deliberation is (1). This section has shown that (2) and (3) are ruled out because, given the causal connections of winning with intending to drink and of intending to win with intentionally drinking, the rules of the puzzle effectively stipulate that one cannot win without intentionally drinking.

In neglecting the link which objective-feasibility constraints provide between intending to win and intentionally drinking, Kavka's decapitation of rational actions from rational intentions renders (the belief-desire inputs of functionally defined) intentions causally idle. With the connection restored, the intention to drink does not enter into rational deliberation as an independent variable but comes packaged in a plan with drinking the toxin. Kavka found the isolation of grounds for rational intentions from grounds for rational actions puzzling. Our findings resolve this puzzle.

We have not yet shown that it is rational to intend to drink. It is still open to Kavka to argue that even within intentional-causal restrictions on the feasibility set, there are rational reasons why Cindy might not be able to intend to drink the toxin. But having purged debate of idle intentions, we are well situated to substantiate the 'madness' Gauthier perceived in Kavka's position. While Cindy's opting for O_3 exhibits consistent deliberation among feasible options,

we will question whether her deliberation is really *rational* deliberation. Where Gauthier's account yields O_1, Kavka's yields only O_3. Since O_3 is clearly inferior to either of O_1 or O_2, the further consideration that O_1 is inferior to O_2 seems less than germane.

Conclusion:

In order to sustain the toxin puzzle, Kavka relied on a mistaken conception of rational agency. Our repairs safeguard Gauthier's dispositions against the objective inability aspects of rational irrationality. Cindy's following the plan of drinking the toxin having received the prize is analogous to Q's practising constraint in the farmers' game. Functionally interpreted, the metaphysics of rational agency does not interfere with an agent's adopting a strategic disposition. To adopt a disposition is simply to follow a dynamically consistent rational plan across a defined range of feasible options.

The decision theoretic apparatus emphasizes how Cindy's and Q's dispositions require warrant from antecedent rationality commitments. To object that Kavka's Cindy chooses a dynamically consistent rational plan, one must assert that Gauthier's dispositions bring unwarranted rationality commitments to bear on the compliance problem. In arguing that it is rational to intend to drink, we have yet to overcome the 'normative coherence' aspects of rational irrationality.

Section 3 Substantiating Overall-Utility Maximization

Even granted causal-intentional restrictions on feasible plans, one could perhaps defend an agent's rational inability to adopt internal constraints on the basis of antecedent rationality principles to which the agent is committed. To squelch this move, we prescribe a rationality principle which requires overall- rather than discrete-utility maximization.

Subsection one shows that an agent engaged in intrapersonal conflict might be rationally unable to commit to a given internal constraint. Subsection two reveals that if the intrapersonal conflict stems from myopically inconsistent intertemporal preferences it can be resolved by governing discrete preferences with a strategically consistent plan. Subsection three worries that if myopia results from temporally biased preferences then overall plans might not be prioritized over the discrete. Subsection four develops an intuitive principle of overall-utility maximization to govern the rationality of temporal biases. Subsection five declares ambivalence concerning a possible formal representation of overall-utility maximization with the standard rational choice principle of independence. Our ambivalence sets up a crucial discussion in subsections six and seven of pragmatic normative justification for principles of rational choice. The pragmatic grounds of normative justification seal our defence of rationally intending to drink the toxin and, by analogy, rationally adopting an internal disposition of strategic constraint.

3.1 Though without explicitly tying the notion to his toxin puzzle, Kavka has argued that rational agents can host internal dilemmas between their constituent subagents.[52] For example, an agent buying a car might value both style and safety. The agent's decision might then be represented as an internal dilemma according to the following matrix (where numbers represent

the order-of-preference of outcomes to that agent rather than representing payoffs to the agent).

Figure 3.1 Intrapersonal Dilemmas

Safety \ Style	Hold Out	Give In
Hold Out	3,3	1,4
Give In	4,1	2,2

1=Most Preferred Model
4=Least Preferred Model

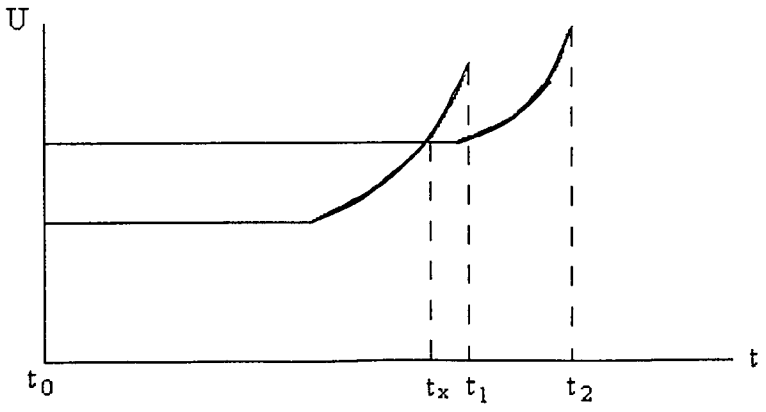
Because subagent 1 will not sacrifice safety for style, and because subagent 2 will not sacrifice style for safety, both hold out. Since there is a value conflict within the agent, she chooses a suboptimal outcome. Kavka does not argue that agents ever actually are comprised of numerous subagents engaged in intrapersonal struggle; internal dilemmas are heuristics for illuminating the nature of value conflicts.

With an 'internal subagent model' of decision making, we could say that Gauthier's solution to the farmers' game results from each farmer's constituent subagents achieving intrapersonal agreements which then enable interpersonal cooperation. Similarly, we could say that an agent's intending to drink the toxin results from the agent's constituent subagents achieving an intrapersonal agreement. Holding a preference for intending to drink the toxin and a preference for not intentionally drinking could then embroil the agent in an intractable internal conflict. And just as it could then be impossible for Cindy to intend to drink the toxin, it could be impossible for Q to precommit to cooperation in the farmers' game. In both cases the agents could be rationally unable to perform the necessary internal manoeuvres.

Internal dilemmas could be easily solved if there were some sort of director-self controlling subservient doer-selves.[53] From the perspective of non-Cartesian metaphysics of mind, resorting to a director Self is of course unpalatable. However directors and doers could be interpreted simply as antecedent rationality conditions which require rational agents to maximize either overall- or discrete-utility respectively. If there is a rationality condition which orders options on the basis of their contributing to overall- or discrete-utility, then intrapersonal conflicts of the sort posed by strategic dispositions will permit solutions on which to base strategic internal constraint. We shall defend just such a rationality condition below.

3.2 There is a growing literature dealing with problems of intertemporal rationality which develops the nature of discrete- versus overall-utility constraints. If an agent arbitrates now between future prospects, then in the time leading up to the payoffs the agent holds a set of intertemporal preferences. A set of preferences is irrationally held if the set is inconsistent. A standard case of inconsistent intertemporal preferences is represented in figure 3.2.[54]

Figure 3.2 Inconsistent Intertemporal Preferences



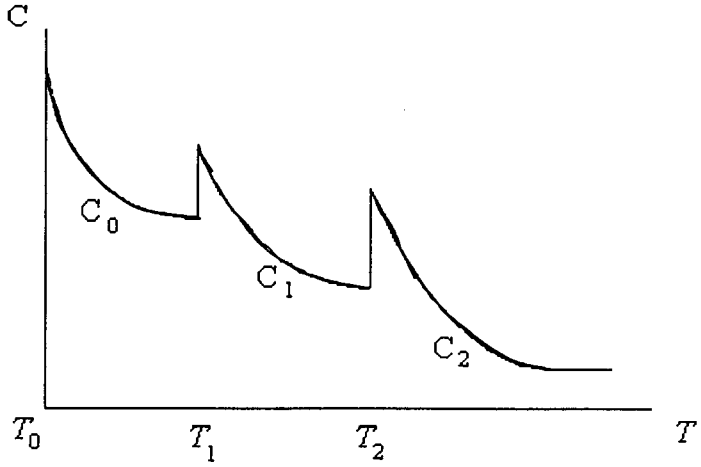
An agent at t_0 is facing a reward at t_2 or a sooner but lesser reward at t_1. At t_0, the agent prefers

the later greater reward. At t_1, she prefers to take the immediate reward rather than to wait for the later payoff. Her preference functions cross at some time t_x before t_1.

At t_x, the agent could be caught in the sort of internal conflict which Kavka anticipated with his intrapersonal prisoner's dilemma.[55]

This agent's inner conflict would be irrational since it results from inconsistent present-value preferences. The irrationality underlying inconsistent preference discounting was emphasized early in the literature by Robert Strotz.[56] In the context of consumption over time and subject to budget constraint, Strotz worried that if a utility-maximizing individual is free to reconsider a rationally chosen consumption plan at some later time, then even when his original expectations of future desires and means of consumption obtain, the agent's future utility-maximizing behaviour might be inconsistent with the consumption plan originally preferred.

Figure 3.3 Myopic Consumption Discounting



Let an agent periodically choose among consumption plans at time T_t, and let the utility-curve C_t be the consumption plan an individual would follow from that time. If an agent does not reevaluate his consumption plan during some temporal period $T_0 < T_t$, he continues to follow C_0. But let the agent recognize at some later T_1 that the stock for consumption available to him has

decreased by an amount consumed from T_0 to T_1 such that consumption plan C_0 is no longer preferred. He then chooses a new best consumption plan C_1. Figure 3.3 represents a series of reevaluations. Since lesser consumption is apportioned to the temporally distant, an agent with the present-value reward functions in figure 3.2 might be interpreted as having the consumption functions in figure 3.3. Strotz argued that agents with such discount functions are guilty of dynamically inconsistent planning. Agents who plan inconsistently make 'myopic' choices. In the parlance of consumption, myopia tags spendthrifts. Inconsistent intertemporal preferences lead agents to making the irrationally myopic choices of a spendthrift.

Strotz allowed spendthrifts two alternatives for rectifying myopic deliberations with 'sophisticated' choice -- strategies of precommitment and strategies of consistent planning. Strotz' precommitments are external devices which enable agents either to commit themselves irrevocably to some action(s) in the future presently deemed appropriate, or else to punish themselves should they misbehave. The upshot of any given precommitment is to preclude grossly unequal allocation of goods within a future period of time when that period moves into the present. This same result might be achieved by internal strategies of consistent planning. The problem is then to find the best plan among those that one can actually follow.

Strotz posed the 'harmony case' to guide planning. Harmony obtains when an agent would apportion the same consumption between T_2 and T_3 from the perspective at T_1 as he would apportion between T_2 and T_3 from the perspective at T_2. That is, harmony obtains when an agent discounts future utilities at a constant rate of interest. An agent achieves harmony by, at each period, selecting his consumption for that period and for the next and allocating this amount between the two periods. In effect, a consistent strategic planner decides to act in any discrete

instance as if he had consistent intertemporal preferences even when he does not. That is, a consistent strategic planner guides discrete-consumption preferences by an overall-consumption plan.

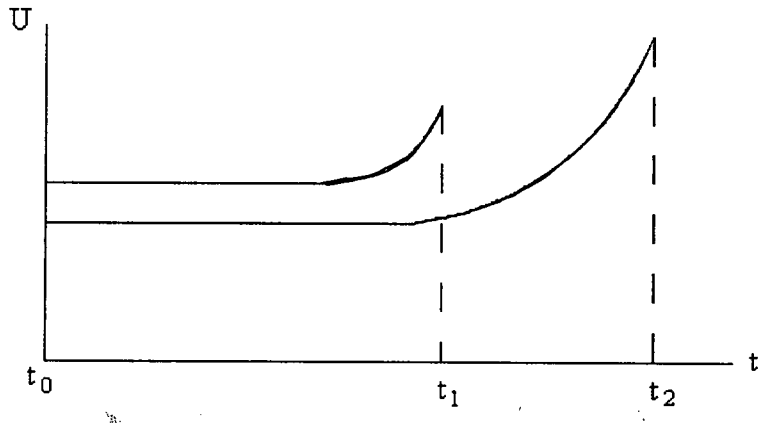
The general attempt to govern intertemporal preferences by formalizable strategies of rational planning appears hopeful for an account of strategic dispositions. A strategy of choosing to follow harmonious plans parallels the notion that constrained maximization must demand real constraint on preferences rather than simply providing straightforward maximization with a disguise. Moreover, the interpretation of consistent planning in light of intertemporal preferences develops the formal decision-theoretic concept of rational agents consistently following feasible plans for action at later times. To choose that plan is to adopt a disposition of strategic constraint.

3.3 Still, these steps are importantly inadequate. Myopia need not be time-preferentially inconsistent, and time-preferentially consistent myopic choices are not self-evidently irrational. To understand these problems, note that stipulating Strotz' harmonious discount rate as a condition of rational planning requires more than the dynamic consistency of preferences; Strotz' argument tacitly presumes the further condition that a rational agent must not bias intertemporal preferences for the near-future or for the present to the detriment of the distant future. This problematic tacit assumption emerges when we recognize that a rational agent could consistently opt for the lesser earlier reward.

Strotz' spendthrift in figure 3.3 does not have to follow the time discounting exhibited in figure 3.2 since the 'curve-jumping' of figure 3.3 need not stem from the curve intersections

in figure 3.2. In figure 3.4 below, an agent exhibits consistent discounting in present-value preference functions (his present-value curves do not cross). But although he will value the reward of t_2 more highly at t_2 than he will value the reward of t_1 at t_1, at t_1 he nevertheless values the reward of t_1 more highly than he values the reward of t_2 from the vantage of t_1.

Figure 3.4 Consistent Temporal Bias



At t_1, the agent consistently prefers taking a smaller immediate reward to waiting for the greater reward. It does not matter that it is easy to manufacture an opposite situation where the agent consistently prefers the later reward over the earlier. The case in figure 3.4 sufficiently suggests that when present-value discounting is consistent, the most we may conclude is that the agent's preferences are temporally biased.

That myopic choice might be time-preferentially consistent suggests that myopia might not necessarily be irrational. An agent's choosing a lesser earlier reward is analogous to Cindy's not-intending to drink the toxin and to Q's not-adopting a disposition of constraint. Cindy and Q might have preferences like those exhibited in figure 3.4. Kavka could press that in the toxin puzzle we face a consistent bias toward the temporally near. Therefore we cannot assume inconsistency of intertemporal preferences simply on the basis of inabilities to commit to internal

constraints.

Let us quickly summarize our progress. So far we have found that an agent engaged in intrapersonal conflict could be unable to commit to a given internal constraint. If the intrapersonal conflict stems from competing intertemporal preferences, it can be resolved by prioritizing one temporal consideration over the other. But if the resolution is myopic as a result of temporal biases, there is no guarantee that the long-term consideration will be prioritized over the short term. A temporal-preference interpretation of dynamic deliberation, one could object, therefore lacks sufficient clout for prioritizing competing intrapersonal concerns. To establish the irrationality of a myopia leading to intrapersonal paralysis, we must move to richer ground.

3.4 Recall from section two that Cindy must deliberate between receiving the prize having drunk the toxin or not-receiving the prize having not-drunk the toxin. Kavka in effect argues that Cindy is rationally resigned to the latter. But since preserving the status quo is (by hypothesis) inferior to drinking the toxin and receiving the prize, and since winning the prize without drinking the toxin is (by the argument of section two) not a feasible option, Cindy can rationally refrain from intending to drink the toxin only if she brings an antecedent commitment to bear on the toxin puzzle. From section 3.3 we learned that the required rationality constraint would have to bias the agent in favour of discrete-utility rather than overall-utility. This subsection develops a rationality principle, a principle of overall-utility maximization, which curtails just such time-preferential biases.

Sidgwick argued that rationality implies impartial concern for all parts of an individual's life. He presented this claim with an analogy between the cumulative parts of an individual's life

and some collective or universal or aggregate 'good' derived from the accumulation of many persons' goods. Just as a difference of time between generations does not itself justify treating different generations differently, neither should a difference in time between different stages of one person's life justify that person's treating the different stages of his life differently. Unlike Sidgwick, Rawls (wisely) does not assume that intergenerational principles are direct extensions of principles of rational choice for one person. However Rawls does agree at the level of the individual that a pure time preference is irrational if the individual does not view all moments as equally parts of his life.[57]

Against Sidgwick-Rawls impartiality, we can set Derek Parfit's argument that at least some temporal biases might be rationally warranted on the basis of forward-looking versus backward-looking derivations of utility.[58] If an agent fears disutility as a result of pain suffered prior to the present even though there are no lingering disabilities, we would likely say that his fears are irrational. That is, one could argue that because a pain is past it ought not generate any present disutility. On the other hand, if an agent fears a pain he knows he will experience in the future, we would likely say that the agent's fears are rational. That is, one could argue that because a pain is still to come it might generate present disutility. If temporal biases are irrational, then a future pain should rationally allow no more present disutility than should a past pain. This result strikes Parfit as absurd. Parfit concludes that some temporal biases are rational.

Parfit's examples superficially complement Elster's more cogent generalizations that an agent unappreciative of the full scope of his life lives only in the present, and that such an agent will consequently succumb to myopia simply because he feels no effects of 'temporal externalities.'[59] Elster notes, where Parfit did not, that temporal externalities could work either

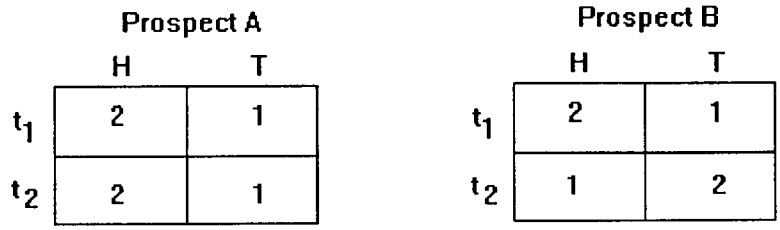
backward or forward and could engender either utility or disutility. Backward externalities obtain when an agent derives present (dis)utilities as a result of (dis)utilities accrued in past experience. Pain might be derived from the cessation of a pleasant experience or pleasure derived from the cessation of pain. Forward externalities include such considerations as the dread or anticipation felt in the present at the prospect of some future event. Let us term forward or backward externalities 'directional' and near-event or distant-event biases 'proximal.'

Given that the past is irreversible, it is irrational for an agent facing a decision to consider all times of her life impartially rather than adopting a directional bias. The plausible rationality of directional biases for rational choice stems from the objective fact that to assess present utilities across past events would be to assess present utilities across objectively non-feasible outcomes.[60] However the bias of the spendthrift is proximal. The rationality of proximal biases does not follow from the rationality of directional biases. The rationality of a proximal bias is a function of the content rather than of the form of the biased events. It might be rational to bias the present at the expense of the future if it means taking a mortgage to buy a house. It might be rational to bias the future at the expense of the present if it means not spending on a day at the races in order to pay off one's mortgage.

Given the relevance of externalities to rational choice, Sidgwick-Rawls impartiality fails in two respects. Neutrality must be violated to satisfy the rationality of a directional bias for the future required by feasibility constraints. Secondly, even a forward-looking impartiality ought not imply that each stage of the agent's future should be equally utility-maximized.

Now consider an agent with two times in his life facing the life-prospects represented in figure 3.5 below.[61]

Figure 3.5 Temporal Prospects



By flipping a fair coin to choose one of two possible lives, the agent has equal chances of living in state Heads or state Tails. By prospect A, his possible life yields 1 unit of good in state H for each of times t_1 and t_2, or 2 in state T for each of t_1 and t_2. By prospect B, his possible life will feature 1 good at t_1 and 2 at t_2 in state H, or 2 at t_1 and 1 at t_2 in state T.

Since the total potential goods at t_1 in A equals the total potential goods at t_1 in B, and since the total potential goods at t_2 in A equal those at t_2 in B, impartiality could be taken to imply that prospects A and B are equally good for the agent. One might object, following Broome, that Prospect B is plausibly better than prospect A since B guarantees three units of good whereas A might yield only two. On the other hand, we could perhaps also object that prospect A might plausibly be defended over B since there is a possibility of getting four rather than three. Either way, there could be reasons not to prefer prospects by impartially weighting the payoffs at individual time periods. That is, based on overall life prospects, the agent could hope for one prospect over the other.

However rational temporal-partiality still does not afford a blanket warrant to either a near or a distant proximal bias. Rational partiality regulates proximal considerations by overall-utility. We clarify with another example raised by Broome.[62]

Figure 3.6 The Salary Game

	Don't Play
Week 1	$200
Week 2	$200

	Play	
	H	T
Week 1	$20	$420
Week 2	$420	$20

Suppose that you will work at a job for two weeks. Your employer offers you a $20 premium each week to toss a coin for 'double or nothing' on your salary. Your employer further guarantees that if and only if you lose in one week you will win in the other. If you decide not to play the game, you get the regular salary of $200 a week without the gaming premium.

Whether you decide to play or not will depend on the extent to which you value an even disbursal of your income. If you want your income evenly disbursed, perhaps in order to support your habit or to buy groceries, you will prefer not to play and will take the $200 a week.

On the other hand, if you have no contingent circumstances which dictate your preference for an even distribution, you do well to play because playing scores a bonus $40 over the two week period. That is, absent context laden externalities, if you decide to play it would make no difference whether the earlier or the later time period afforded a greater payoff. A proximal temporal preference which interfered with your ability to commit to playing would be irrational since the bias against waiting for cash would deny you a feasible bonus. That the greater payoff might obtain later rather than earlier (or earlier rather than later) does not alter the rationality of playing rather than not-playing.

We must generalize the externality-laden case and the externality-free case under a

unified principle of rational choice. The principle must importantly presume that an agent's existence consists in a series of temporal subagents constituting a unified life. The goodness of an agent's life will consist in an aggregation of goods accumulated by the temporal subagents. The principle must be sensitive to contextually rich externalities. Furthermore, the principle must allow that the proximity of an outcome could be relevant to rational choice to the extent that proximity affects the probability of receiving or not receiving payoffs.[63] Loosely stated, a rational agent must tally all (dis)utilities predictably accrued into a sum good of his overall life before deciding whether to weight prospects at each temporal stage as more or less important than prospects at other stages. Proximally-biased preferences will then be made rational by the extent to which they further or hinder the agent's overall-utility maximization.

Thus we arrive at a crude principle which constitutes a robust antecedent constraint with which to regulate temporal biases. With the qualifications made for directional biases and for externalities, the overriding constraint is that an agent acts rationally in a decision situation if and only if acting so as to maximize (expected) overall-utility.

In the toxin puzzle Cindy is guaranteed a later greater reward if she forgoes a lesser earlier prize. If she does not treat her life as a whole, she is not able to intend to drink the toxin but ends up living an inferior life. By the principle of overall-utility maximization, an agent who is unable to intend to drink the toxin on the basis of preferences proximally biased against the distant future is an irrational agent. The explanation of this irrationality lies in the structure of dynamic choice. Particular decision problems will become subordinated to a broader framework of rational life plans. Since intending to drink the toxin affords greater overall-utility to Cindy's continuing life than does not-intending to drink the toxin, then by the dynamics of

intertemporal utility maximization she is rational to intend to drink the toxin even if at the given instance of her life she must thereby incur disutility.

3.5 In so far as it restricts the content of rational preferences, a principle of overall-utility maximization is a substantive rationality constraint. At first blush, substantivism conflicts with the sop we extended to value-subjectivity in section 1.1. If rationality really only concerns the means to subjectively determined ends, then we ought not to be able to decide that some ends are more rationally pursued than others. The standard response to such an objection argues that no such tension obtains when substantive restrictions are applied to patterns of preferences rather than to the content of individual preferences. If, for example, an agent's preferences are inconsistent, then even if no individual preference is by itself irrational, the set could still be an irrationally held set of preferences.

To argue such points, rational-decision theorists usually abstract substantive principles of rational choice from contextually defined cases and analyze consistency requirements formally. In this vein, Broome hints that the restraint we require on sets of preferences might be provided by the principle of independence.[64] Under one common articulation, the independence condition obtains when for any three gambles g_1, g_2, and g_3, g_1 is preferred to g_2 just in case a lottery between g_1 at probability p or g_3 at probability 1-p is preferred to a lottery between g_2 at the same probability p or g_3 at probability 1-p.[65] Independence encapsulates the intuition that an option x be preferred to an option y just in case x is ordered ahead of y regardless of whatever other options might or might not be added to or excluded from the set.

To see how a principle of overall-utility might be built upon independence, we turn to

rudimentary set theory.[66] For some set $X=\{x_1, x_2,...x_n\}$, we can define a *vector* $(x_i, x_j,...x_n)$ as a listing comprised of n *locations*, each location filled by an *occurrence* from the field of X. Let us further say that some relation R obtains in X just in case for any two alternatives x_i and x_j, x_iRx_j just in case x_i is preferred or is indifferent to x_j. (Let us temporarily leave the criteria determining the ordering of preferences by R undefined.) X is said to be *connected* just in case for every x_i and x_j in X, either x_iRx_j or x_jRx_i. X is said to be *transitive* just in case for any triplet x_i, x_j, and x_q, if x_iRx_j and x_jRx_q then x_iRx_q. If X is transitive and connected, then R is said to define a *weak ordering* (of the members of) X. For X to be a strong ordering it would have to be the case that R obtains if and only if for any two alternatives x_i and x_j, x_iRx_j just in case x_i is strictly preferred to x_j.

We can construct two subvectors of X by twice altering occurrences, say at locations x_1 and x_3, while keeping the occurrences at all other locations constant. That is, we can choose two subsets of X, $X(S_1)=(x^*_1, x^*_3)$ and $X(S_2)=(x^\#_1,x^\#_3)$. It is standard to say that $X(S_1)$ is conditionally ordered ahead of $X(S_2)$ by relation R just in case $(x^*_1, x_2, x^*_3, x_4,...,x_n)R(x^\#_1, x_2, x^\#_3, x_4,...,x_n)$. A subset of locations is said to be separable from X under R if and only if the conditional ordering across all the vectors corresponding to the subset remains the same regardless of the occurrences at other locations. That is, some ordering (x_i, x_j) is independently ordered just in case x_iRx_j even when no other occurrences at the locations in X are held constant.

Now assume that X is a weak ordering of options x_i. X is said to be *strongly separable* just in case every subvector of locations (i.e. every individual location, every doublet of locations, every triplet of locations,...) is separable. And now let us declare that the criteria warranting an ordering of preferences by R is the utility afforded an agent from each of the

possible outcomes. X is then *additively separable* just in case it can be represented by a sum of utilities $U(X)=u_1(x_1)+u_2(x_2)+...+u_n(x_n)$. An additively separable ordering salsifies the condition of independence. Moreover, if the options $x_1, x_2,...x_n$ do represent utilities derived from possible outcomes, X then represents the agent's expected utility function. If X is additively separable, then we can calculate the expected overall-utility afforded by various orderings of the outcomes.

In the salary game (figure 3.6), the principle of overall-utility declares that an (externality-free) agent ought to prefer playing to not-playing no matter whether the payout comes in stage 1 or stage 2. That is, the agent should prefer playing to not-playing on the basis of an expected overall-utility derived from summing the possible payoffs afforded in each period by the given strategy. We might say that the ordering '(playing)R(not-playing)' is strongly separable, and, by independence, that playing is preferable to not-playing. It could therefore appear that in requiring overall-utility maximization, we have employed independence as a condition of rational choice. We must resist this characterization of our argument.

Although independence is often thought to be a formal consistency condition for rational sets of preferences, independence is too strong a condition to impose on rational choice. Independence assumes that outcomes in some state are evaluated independently of outcomes in another state. Following Elster, however, we have pinned our sense of overall-utility to an agent's continuing existence or 'full scope of life.' Moreover, in section two we introduced natural causal mechanisms of rational agency as relevant filters on decision making. Thus in defending a condition of overall-utility maximization, we cannot evaluate outcomes independently but must sometimes evaluate outcomes in one state on the basis of outcomes in other states.

For example, let X be the set of possible outcomes in a bounded dynamic decision problem like those in figures 2.1 and 2.3. Each vector of X is comprised of locations represented by the possible plans within that decision problem. Occurrences which fill the locations are represented by truncations which structure possible plans. As argued in section two, causal relations could govern the ordering of occurrences at different locations in X. Following the jargon, these causal relations between locations could conceivably count as peculiar *complementarities* between occurrences in a feasible set. (Complementarities are factors which influence the orderings of outcomes by virtue of reasons that do not strictly stem from, or that potentially interfere with, an ordering of preferences by R.) Since it takes account of the natural mechanisms of agency, our principles of overall-utility might not then require that the locations and occurrences of X be ordered independently. That is, overall-utility maximization could be plausibly interpreted as violating the independence principle.

We should not yet conclude either that a principle of overall-utility maximization does or does not violate the condition of independence. Rather, we do better to notice only that formal characterizations of choice are too easily manipulated to adequately carry the burden of argument. Indeed, we can be ambivalent about violating independence since there is a difference between formal conditions on sets of preferences and normative conditions on rational choice. The next subsection substantiates this crucial distinction.

3.6 Though traditionally unquestioned as basic consistency requirements on the preference sets of rational choice, independence and transitivity conditions are becoming increasingly contested. If our principle is not built upon independence and transitivity, perhaps that is so

much the better for our principle. We are concerned that overall-utility be justified as a principle of rational choice. If we can understand how consistency, broadly construed, has become a lynch pin of rational justification, we will have the material needed to cap a defense of overall-utility maximization as a condition on rational choice without relying on ambiguous formal conditions.

We begin with arguments raised by Frederic Schick against standard justifications of coherent probability assignments.[67] Suppose you face three possible bets, and that you would be willing to pay x for the first, y for the second, or z for the third. But suppose as well that if you were to take a book of the three bets at those rates and win, the total winnings would be less than the price you paid for the three (x+y+z). Clearly, even if each of the three bets on its own stands to increase your utility, if you take the whole book you take a bad bet. You have been 'Dutch booked.'

Dutch book troubles have been standardly thought to result from an agent's incoherent valuations of probabilities.[68] Let there be two mutually exclusive events h and k, where I expect h to obtain with probability p(h), k with probability p(k), and h or k with p(hvk). (The 'v' in 'hvk' operates as a disjunctive 'or'.) For stakes S I am then willing to pay p(h)S for the bet that yields S if h but otherwise nothing, p(k)S for the bet that yields S if k but otherwise nothing, and -p(hvk)S for the bet that yields -S if hvk but otherwise nothing. (To pay -x means demanding a payoff of no less than x; a gain of -x is to have lost x.) Suppose I buy the three bets together. Either h obtains, k obtains, or neither. If h, then I win both p(h)S and -S. If k, then I win both p(k)S and -S. In each case the total winnings sum to zero. If it is not the case that (hvk), I win no bet and gain zero. Taken together, the three bets then yield nothing. In taking the three bets I will be Dutch booked unless p(h)+p(k)=p(hvk). The sum of the prices, (p(h)+p(k)-p(hvk))S,

would be more than zero only where p(hvk)<p(h)+p(k). Knowing this, I can be Dutch booked only if I incoherently valuate probabilities for the events across which I am betting.

In order for this demonstration of the incoherence of Dutch booked probabilities to hold, it must be assumed that I am willing to pay (p(h)+p(k)-p(hvk))S for the three together. This assumption is just the principle of additivity. For our purposes, it is interesting that additivity in Dutch book cases rests on independent valuations of the events. In the Dutch book context, independence requires that where I know my betting portfolio, the values I set on possible new bets will remain unaffected by this knowledge. For a Dutch book argument to have bearing on a theory of rationality as expected utility maximization, it must be assumed that the worth an agent assigns some event A remains constant regardless of whether or not she thinks some further event B is also in effect, and likewise with valuations of event B and some further event C.

Dutch-books are cousins to money pumps. Imagine that an agent prefers apples to oranges, that she has an orange, and that you have an apple. It is plausible that she pay you some modest amount to exchange her orange for your apple. This transaction completed, suppose you next reveal that you also have a pear. Preferring pears to apples, the agent then offers you a small amount to exchange her newly acquired apple for your pear. The transaction complete, the agent then confesses that she really prefers oranges to pears and suggests that she pay you a modest sum to exchange your orange for her pear. You are of course delighted since exchanging for the orange sets up another apple exchange, another pear exchange, another orange exchange, another apple exchange, and so on. Since your friend pays you a modest fee to make each exchange, you have found a 'money pump.' Money pumps are engendered, it is

standardly argued, when preferences are intransitive.

The money pump argument for a principle of transitivity evidently assumes that where a series of transactions has been made, the value the agent sets on the next in a series is the same as the value that would have been set on the transaction had it not been one in a series. That is, as with the Dutch-book case, there is again an assumption of (intertemporal) additivity built into the principle of transitivity. The additivity assumption presumes that each of a series of transactions is value-independent.

Suppose we assume intransitive preferences and let the total gain for some sequence of outcomes be less than the sum of what an agent would pay for each of the events singly. All that follows is either that the preferences are incoherent or that the values set on each event are not set independently. We could do the same with the probability assignments in Dutch-books. It does seem that Dutch bookable probability assessments and money pumpable preferences are both 'exploitable' and incoherent. But their incoherence, Schick concludes, stems not from their exploitability but either from their inconsistency or from their violation of independence.

Schick's discussion challenges the use of Dutch books and money pumps to justify the coherence probability assessments and the transitivity of preferences as formal conditions of rational sets of preferences. We can explicate the challenge with a distinction between descriptive and normative accounts of rational choice. The rational choice formalism has primarily been tested and prized for its descriptive successes. Suppose we assume that an agent is a rational agent, and then manipulate a formalism to derive predictions about that agent's behaviour. Suppose further that our predictions become instantiated. We can then be confident in the explanatory powers of our formalism. This procedure forms the basis for descriptive

justification (d-justification) of a formal principle of rational choice.

Schick's challenge seems like a challenge from the point of view of d-justification. Schick essentially argues that we cannot be confident in transitivity as a necessary feature of rational choice since all that follows from an observation of intransitivity is either that the preferences are incoherent or that the preferences violate independence. And if there is good (descriptive) reason not to assume that independence holds, it becomes difficult to infer that a money-pump argument is enough to show that rational preferences must be transitive.

Such limits on d-justified principles potentially forge a gap between d-justification and normative justification (n-justification). Since von Neumann and Morgenstern, it has been widely accepted that in order to define an agent's utility-function, the agent's preferences need at least to satisfy transitivity and connectedness (if not completeness or monotonicity). But except in a trivial sense of 'rational' where x is rational just in case x is described by the formalism of rational choice, it is importantly not the case that an agent is called rational only because her preferences satisfy certain formal conditions. We do better to say that certain conditions are needed in order to formally represent an agent's rationality. That is to say, there might be aspects of rationality that are not yet captured in the formalism. The point of n-justification is that we describe the choices which satisfy our formal requirements as 'rational choices' only if our formalism satisfactorily represents some further normative content.

In the cases of Dutch-books and money pumps, let us grant Schick's conclusion that either the preferences are incoherent or that the preferences violate independence. It still makes sense to ask what could make intransitive non-independent preferences irrational. It would seem that they are irrational because they are exploitable, not because they are incoherent. They are

irrational because even if they could be made 'coherently intransitive,' the agent would do worse if she encountered an exploiter than she would do if her preferences were transitive. Intransitive preferences would be irrational not on some strictly formal inconsistency property but by some rather more ambiguous property of being exploitable. We then normatively endorse transitivity and the coherent valuation of probabilities not because each serves as a descriptively successful formal property but because the formal properties are the best approximations we have of non-exploitability.

3.7 A sceptic might object that in posing non-exploitability as a key to n-justification, we have either slipped into a discussion of imperfect rationality or we are begging the question of an agent's continuing life.

The objection from 'imperfect' rationality is surely specious. In his characterization of the foundations of rational justification, McClennen notes that since norms prescribe rather than describe behaviour they are not hypotheses subject to empirical tests.[69] McClennen's generalization is not quite precise. Above we considered the possibility of an agent proximally-biased to the near. We found that the agent was an irrational agent because she could do better for herself if she was partial to overall-utility rather than only to near-utilities. More generally, the efficacy of a norm for promoting the agent's own ends constitutes an empirical test for normative principles of rational choice.

Empirical tests of rational norms can be coupled with Herbert Simon's notions of bounded rationality.[70] Simon essentially argued that agents which are metaphysically limited in their abilities to gather and process data can actually maximize their utility by following cost-

effective rules of thumb (or 'satisficing'). That is, where excessive deliberation actually brings about disutility, agents do better for themselves in the long run by satisficing even though they might on occasion perform acts which yield inferior results to other acts possible on that occasion. One must not object that Simon's bounded rational agents are imperfectly rational.[71] When the rules of thumb are indeed cost effective, the agents which follow these rules do better for themselves than do the agents which deliberate on each occasion. That is, these agents employ the more effective means to their ends. And by definition, 'perfect' rationality is simply the choice of the best means to given ends. An objection that agents who act in light of their circumstances in the natural causal world are imperfectly rational agents likewise misses the mark.

The more compelling objection is that we have simply assumed that an agent's ends and an agent's own standards are defined and achieved over the course of a life rather than across decision instances. There are three answers to this objection. Firstly, by (crude) definition, to achieve utility is for an agent to get what the agent wants. Rational choice theory need not assume any metric of utility that all agents must maximize; rational choice theory need only assume that each agent maximizes some metric of utility. Normativity is then driven by the agent's own standards. Furthermore, recall that agents act and choose in a world of natural mechanisms. Any agent that systematically opts for a lesser nearer utility over a greater later utility will, given time, likely cease to be an agent in the natural causal world. Failing demise, the agent will certainly acquire less of what it wants than what it can get.

There are of course some exceptional circumstances which could sustain the objection. We could conceive of agents who prefer to be poor or even to cease to exist. These

circumstances motivate our second response. The objection that she might choose to be impoverished is not an objection to overall-utility maximization. The measure of overall-utility is not net-gain but utility. If an agent wants to be impoverished, we could say that her abdication of 'real terms' dividends promotes her utility. An agent with myopic norms chooses lesser utility, not lesser net gains. An agent who consistently chooses lesser utility does worse by her own standards than she would do if she opted for greater utility even if her standards require her to seek poverty. Moreover, although the rationality of suicide poses an interesting question, it is clearly at most a qualifying case rather than a case from which to generalize principles of rational choice.

The objector could press that we can also conceive (and build) 'worlds' which consist of just one time. The notion of overall-utility then becomes unintelligible at most and at least superfluous. At this point, the sceptic's objection reveals itself as plainly silly. When rational choice theory attempts to test and refine principles of choice for situations without plausible analogues in a world even vaguely like ours, the critical thrust of normativity in rational choice theory is lost. We pursue rational choice theories precisely because they have proven powerfully explanatory of phenomena in our very own world.[72] A theory which cannot support developments from 'one-instant' test cases to test cases which involve an agent's enduring life is an uninteresting theory. That agents have 'lives' that must be protected from exploitation is not a question begging assumption; it is the reason for asking any questions at all.

3.8 Such an account of n-justification is unabashedly pragmatic. The pragmatism provides an important consideration otherwise missing from rational justification.

Consider the toxin puzzle once again. In order that it be irrational for an agent to intend to drink the toxin, the agent must be committed to discrete-utility maximization. We have argued that discrete-utility maximization is irrational. Discrete-utility maximization is in effect the handling of one of a series of events as if it were value-independent from any earlier or later events. Thus, we could plausibly suppose that Kavka's argument runs from an assumption of independence.

Our own principle of utility-maximization could be plausibly construed as violating independence. If such an interpretation holds, it still provides insufficient grounds for resolving the debate between Kavka and Gauthier. In order to evaluate the rationality of violating independence, we had to invoke 'deeper' notions of salient natural mechanisms conducive to an agent's flourishing in her environment. Having unearthed the pragmatism which grounds normative justification, we may invoke that weapon without relying on the further, tenuous formal representations of the relevant principles.

Conclusion:

The implications of the above for Gauthier's strategic dispositions are obvious. The argument in 1.2 for the rationality of constrained maximization over straightforward maximization turned on CMs doing better in their environment than do SMs. A principle rationalizing straightforward maximization would need to rationalize discrete-utlility over overall-utility maximization. Such an argument fails. Without that argument, there can be no antecedent rationality commitment which could tangle either Cindy or Q in an intractable intrapersonal dilemma. Moreover, given the structure of intertemporal decisions as rational

plans, and given the function of causal-intentional structures in rational planning, agents will also be objectively able to adopt dispositions of constraint. By emphasizing the structure of intertemporal rational choice theory, and by filtering this structure through a pragmatic conception of justification, we have indirectly supported a Gauthier-like argument for rational compliance.

Conclusion.

David Gauthier argued that it is rational to adopt strategic dispositions of internal constraint which make it individually rational to comply with a joint strategy even in one-shot Prisoners' Dilemmas. These dispositions take the form of rational metastrategies for rational choice. We defended Gauthier's argument against potential objections from standard accounts of rational choice.

In section one we noted that rational choice is best evaluated instrumentally, and that the options for choices are grounded in primitive beliefs and desires and preferences. From these precepts, it should be clearly evident that the natural mechanisms of agency are central to the structure of rational choice. Thus in section two we argued that if an agent rationally intends to A, and if A causes B, then a rational agent will intentionally B. Extending this argument with an apparatus of rational plans, we further concluded that if it is rational to precommit to constraint, then (other things being equal) it will be rational to act in accord with the precommitment.

The onus for defending the rationality of internal constraints on strategic behaviour thus shifted to the rationality of precommitting to given plans. Section three addressed this concern with the consideration that a rational agent must act in accord with the plan which maximizes overall-utility. Given this condition on rational choice, an agent cannot rationally hold the discrete-utility maximizing disposition necessary to undermine Gauthier's strategic dispositions.

There is a decidedly pragmatic undercurrent permeating this defense of strategic dispositions. Broadly speaking, section three hinted that the principle of independence is too strong a condition on rational choice. The argument was not that independence is unnecessary

for coherent probability valuations or for coherent preference sets or for formally representing utility functions (on these matters we hinted ambivalence). Rather, we suggested only that the formal consistency requirement is plausibly too strong for normative rational choice. In clarifying this claim we drew a crude distinction between normative justification and descriptive justification and suggested that normative content ought not be held hostage to descriptively justified formalism. In effect, we suggested that there could be rational normative content not fully reflected in the technical apparatus.

Our candidate for this role appeared in the guise of rudimentary 'evolutionary' normativity. The guiding tenet here is that the best test for a normatively justified principle is the extent to which it works for the agent(s) that hold the norm. Norms which demonstrably promote an agent's continuing life are better than norms which do not. Further, we intimated that the norms of agents which flourish in their environments are more rational than the norms of agents that do not.

Rather than being subversive ideas, the arguments of sections two and three suggest that this sort of evolutionary criterion for rational evaluation is already grounded in the founding principles of rational choice theory. If rationality is instrumental, then agents which maximize their ends within the limits imposed by their natural endowments are more rational than those that do not. Moreover, since Gauthier's dispositions enable agents to flourish in their environments, Gauthier's dispositions thus exploit the central premises of instrumental rationality. Rational choice theorists should welcome this result.[73]

Notes:

1.Gauthier 1986.

2. Rawls 1971.

3. Rawls 1980, 519.

4. This broad interpretation of Rawls largely agrees with Rorty's sweeping generalizations in Rorty 1991a; Rorty 1991b.

5. Gauthier 1988, 385.

6. Unless authors are specifically noted, the following account of 'standard rational choice apparatus' is gleaned from Brams and Kilgour 1988a; Brams and Kilgour 1988b; Cudd 1993; Elster 1989a; Elster 1989b; Gauthier 1986, Chpts 2-3; Gauthier 1990b; Gauthier 1990c; Hampton 1994; Harsanyi 1965a; Harsanyi 1965b; Harsanyi 1976; Harsanyi 1977; Moser 1990; Rapaport 1966; Savage 1990.

7. I here synthesize Brian Skyrms' paraphrasing of the paradox with Bernoulli's original treatment. Skyrms 1990, 3ff; Bernoulli 1954, 23-36.

8. It is standard to read 'x>y' as 'x is preferred to y.'

9. For example if event x_1 is valued at u_1 with probability p, x_2 at u_2 with probability q, and x_n at u_n with probability 1-p-q-...-r then the value of a lottery across x_1, x_2,...x_n is $[u_1p+u_2q+..u_n(1-p-q-...-r)]$.

10. See von Neumann and Morgenstern 1944. For the mathematically modest, Alvin Roth and John Broome each present clear derivations of the essential von Neumann-Morgenstern utility function. Roth 1979, 2-3; Broome 1991, 65ff.

11. More carefully stated, a strategy is pure if it assigns a probability 1 to one action in the

lottery and a probability zero to all others; a strategy is mixed if it assigns non-zero probabilities to more than one action. A pure strategy has one action as a prize, a mixed strategy has more than one action as prizes in the lottery.

12. Strictly speaking, strategic decision making is the realm of game theory. There is debate whether or not game theory is a distinct branch of rational choice from individual decision theory or if one subsumes the other. Since strategic outcomes simply are the products of individual choices, we blur the distinction between game theory and decision theory.

13. The "Prisoner's Dilemma" comes from a tale of two prisoners being held with weak evidence for a conviction. Each is separately offered reprieve if he unilaterally informs on the other, but each faces a heavy sentence if the other unilaterally informs on him. Both receive moderate sentences if both inform, both get only light sentences if neither informs. Originally attributed to A.W. Tucker, the story was formalized in Luce and Raiffa 1957, 95.

14. For further substantive discussions of compliance problems see Hardin 1982; Olson 1965.

15. Ideas similar to this solution can be found in Sen 1977 and in the discussion of endogenous preference changes in Elster 1984.

16. For insightful discussions of coordination problems and solutions see Thomas Schelling 1963, 89ff; David Gauthier 1975; David Lewis 1969.

17. An internal/external distinction is sometimes taken to characterize a distinction between moral and political constraint. For example Gauthier invokes this distinction when arguing that Hobbes' sovereign is a political, not a moral, solution to compliance (1986, 163). I see no reason to advance this moral/political distinction; morals and politics might well cut across both internal and external constraints.

18. These examples are, respectively, general allusions to Thomas Hobbes' *Leviathan*, Schelling's *Strategy of Conflict*, and David Hume's account of the origin and justice of property rights in his *Treatise of Human Nature* Book III, Part 2, section ii.

19. Notice that Gauthier does not use a farmers' game at this juncture; we formalize it for our purposes from his adaptation of Hume's story in Gauthier 1994.

20. The case of multiple one-shot PDs is an importantly different case from the iterated dilemmas in Robert Axelrod 1981, 306-318. Where finite and infinite iterations both do work for Axelrod, it is only the disposition that does work in this argument of Gauthier's.

21. In this paper we endorse the metastrategic form of dispositions but not necessarily the particular contents. For example, there are easily built societies where we would favour Peter Danielson's Reciprocal Cooperator over Gauthier's Constrained Maximizer. For important critical clarifications along these lines see Danielson 1991; Danielson 1988; Robert Frank 1987.

22. The objection is that of Hume's Sensible Knave; Gauthier 1986, 181.

23. Kavka 1983, 35.

24. By Derek Parfit's treatment, it is not necessarily a damaging implication that a theory rationalizes instances of irrationality; see Parfit 1984, 9-13. Kavka more directly construes the implication of rationalizing irrationality, at least in cases of deterrence and threat behaviour, as offering a *reductio* of constrained maximization; see Kavka 1993a, 350.

25. Kavka 1984, 155-159; Gauthier 1994, 697-702; Gauthier 1984, 159-161.

26. Gauthier further requires that the agent compare these expectations with her expectations of which act(s) would make her life go better had she not performed the intentionally restrictive act. These counter-factual expectations help justify not-honouring foolishly offered assurances. For

this chapter, these points are not immediately germane.

27. Gauthier 1994, 702-707.

28. Kavka 1983, 36; 1984, 157.

29. Gauthier 1994, 709.

30. I am grateful to Paul Bartha for raising this interpretation of Kavka's objection.

31. David Lewis 1984, 141-155.

32. Kavka 1984, 158.

33. Gauthier 1984, 161.

34. See Donald Davidson 1980.

35. For a clear presentation of this standard assumption of rational choice theory without reference to models of agency see Jon Elster 1989, 4ff.

36. A dead giveaway that Kavka generally adopts Davidson's theory of intentional action is found in his explicit claim that he generally adopts Davidson's theory of action; Kavka 1983, 35. However Kavka's response to Lewis reveals significant human-psychologizing extensions to this endorsement which we suspect Davidson would not necessarily follow.

37. For determining intentional agency we would generally endorse the approach advocated in Daniel Dennett 1990.

38. Note that ours is a potentially subversive refinement of Gauthier's position. In responding to Christopher Morris 1988, Gauthier has expressed a growing interest in the metaphysics of Self as relevant to the defence of constrained maximization. Coupled with his characterization of agents as semantic representers, Gauthier's interest in the Self intimates a tendency toward psychological realism which we clearly do not share. See Gauthier 1988.

39. Mele 1992.

40. We scare-quote 'pick' in order to preempt the unhelpful objection that 'choosing to unintentionally *A* is not to unintentionally *A* at all for if it is chosen it is not unintentional.'

41. Mele, 183-4.

42. Michael Bratman 1987.

43. As an example of Bratman's enormously accommodating breadth, consider these as partial, hierarchical plans: "I wish to play basketball sometime tomorrow," "My ambition is only to play basketball in university," "I dream of playing basketball all next summer," or "I'm going to play basketball every day this summer and still finish my thesis."

44. Bratman 28.

45. McClennen 1990. See also Brian Skyrms 1990.

46. This diagram largely reproduces McClennen 1990, 100, Figure 6.1.

47. We take this formal definition from McClennen 1990, 120; note that Peter Hammond 1977 defines dynamic consistency with potentially significant alternative notation. The differences do not affect our interpretive arguments, however, but concern individual's preference-orderings being treated in a manner like to aggregating social welfare functions.

48. Gauthier 1994, 709.

49. By hypothesis, the billionaire's decision will be fully determined and Cindy will know the outcome with certainty given that she knows the sincerity of her own intention. We use chance notation because the move is made in the tree by a player other than Cindy.

50. The packaging obtains only failing some unexpected causal intermediary. In the toxin case, because the agent knows ahead of time that the prize will be won before the time comes to drink the toxin, this prior information cannot serve as a later causal intermediary to prevent drinking

once the intention to drink has been formed. Furthermore, since the decision situation in the morning does not involve any updated information, a Bayesian analysis cannot bolster Kavka's case.

51. There are of course more than three possible combinations of winning and drinking but we don't need to consider the ones that include the combination 'not-getting the prize but drinking' because that is clearly not ever going to be a utility-maximizing option.

52. Kavka 1991; Kavka 1993.

53. We loosely adapt the director-doer vocabulary from Thaler's 'planner-doer' interpretation of Strotz in Thaler 1980.

54. Diagrams of this sort are used for illustrative purposes similar to ours in Parfit 1984, Ainslie 1975.

55. For a detailed argument that crossing present-value curves yield ambivalence see George Ainslie 1992, 60ff. Due to his preoccupation with foibles of human psychology rather than with formal matters of rational choice, we do not wish to relate our use of temporal-preference theory too closely with Ainslie's. Note as well that employing intertemporal preference theory enables us to effectively side-step the traditionally tempting philosophical notion of *akrasia* (or weakness of the will) to explain myopic behaviour. As Jon Elster explains, weakness of the will requires that four conditions obtain: (I) there is a *prima facie* assessment that X is good, (ii) a *prima facie* assessment that Y is good, and (iii) an 'all-things-considered' judgement that X is better than Y; but (iv) Y is nevertheless chosen. Jon Elster 1985, 250. The 'compulsion' to choose Y is already accounted for. If an agent has crossing present-value discount functions due to inconsistent intertemporal preferences, then at the time her preferences cross she will be unable to commit

herself to the long-term benefit. *Akrasia* clearly adds nothing more to debate than a quaint psychological notion and a cool Greek word.

56. Strotz 1955.

57. Sidgwick 1907, 381; Rawls 1971, 295.

58. Parfit 1984a; Parfit 1984b.

59. Elster 1985, 236ff. Also, it is important that Elster and Parfit differ significantly in that Parfit sympathizes with a near-future-bias as not-irrational whereas Elster exploits time-biased preferences to counteract even consistent myopia. See Elster 1984, 65-76.

60. Parfit might worry about the objectivity of this 'fact.' In keeping with the pragmatic methodological tradition (developed below) which underpins a normative theory of rational choice, such a "philosophically interesting" objection is irrelevant.

61. Broome 1992; Broome 1991, 228ff.

62. Broome 1991, 62ff.

63. For example, imagine that it is not irrational to take up smoking late in life because the chances of accrued bad effects catching up to you before you die are slim.

64. Broome 1991, 62.

65. That is, the preference relation g_1Rg_2 satisfies independence just in $case_1(g(p),g(1-p))R(g_2(p),g_3(1-p))$.

66. Our account below synthesizes Broome 1991, Chpts 4- 5 with McClennen 1990, Chpts 2-5.

67. Frederic Schick 1986.

68. There are four standard criteria of coherent probabilities: that a probability sample space equal one, that probabilities be non-negative, that they be additive, and that the probability of event A given event B is equal to p(A/B).

69. McClennen 1990, 60.

70. Simon 1966, Simon and March 1958.

71.cf. March 1978.

72. For impressive results see Olson 1965, Downs 1957, Becker 1976, and those catalogued in Hardin 1982.

73. I am grateful to Peter Danielson and to Paul Bartha for helpful comments on earlier versions of this paper.

References

Ainslie, George. 1975. "Specious Reward." *Psychological Bulletin* 82, 463-496.

---. 1992. *Picoeconomics*. New York: Cambridge University Press.

Axelrod, Robert. 1981. "The Emergence of Cooperation Among Egoists." *American Political Science Review* 75:2, 306-318.

Becker, Gary S. 1976. *The Economic Approach to Human Behaviour*. Chicago: University of Chicago Press.

Bernoulli, Daniel. 1738. "Exposition of a New Theory on the Measurement of Risk." translated by Louise Sommer from "*Specimen Theoria Novae de Mensura Sortis*." *Econometrica 27* (1954) 23-36.

Brams, Steven J. and D.M. Kilgour. 1988a. "National Security Games." *Synthese* 76, 183-200.

---. 1988b. *Game Theory and National Security*. New York: Basil Blackwell.

Bratman, Michael. 1987. *Intention, Plans, and Practical Reason*. Cambridge, Mass: Harvard University Press.

Broome, John. 1991. *Weighing Goods*. Oxford: Basil Blackwell.

---. 1992. "Bernoulli, Harsanyi, and the Principle of Temporal Good." In Selten (1992), 353-373.

Buchanan, James. 1975. *The Limits of Liberty: Between Anarchy and Leviathan*. Chicago: University of Chicago Press.

Cudd, Ann. 1993. "Game Theory and the History of Ideas about Rationality." *Economics and Philosophy* 9:1, 101-133.

Danielson, Peter. 1988. "The Visible Hand of Morality." *Canadian Journal of Philosophy* 18:2, 357-384.

---. 1991. "Closing the Compliance Dilemma: How it's rational to be moral in a Lamarckian world." in Vallentyne (1991), 291-322.

Davidson, Donald. 1980. *Essays on Actions and Events*. New York: Oxford University Press.

Dennett, Daniel. 1990. *Intentional Stance*. Cambridge, Mass: MIT Press.

Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: Harper and Brothers.

Easton, David (editor). 1966. *Varieties of Political Theory*. Englewood Cliffs: Prentice-Hall.

Elster, Jon. 1984. *Ulysses and the Sirens*. Revised edition. Cambridge: Cambridge University Press.

---. 1985. "Weakness of the Will and the Free-rider Problem" *Economics and Philosophy* 1, 231-265.

---. 1989a. *Solomonic Judgements.* New York: Cambridge University Press.

---. 1989b. *Nuts and Bolts for the Social Sciences*. New York: Cambridge University Press.

Frank, Robert. 1987. "If *Homo Economicus* Could Choose His Own Utility Function, Would He Want One with a Conscience?" *American Economic Review* 77:4, 593-604.

Gibbard, Alan. 1990. *Wise Choices, Apt Feelings*. Cambridge, Mass: Harvard University Press.

David Gauthier. 1975. "Coordination." *Dialogue* 14:2, 195-221.

---. 1979. "David Hume, Contractarian." *The Philosophical Review* 88:1, 3-38.

---. 1984. "Afterthoughts." in MacLean (1984), 159-161.

---. 1986. *Morals By Agreement*. New York: Oxford University Press.

---. 1988a. "Morality, Rational Choice, and Semantic Representation." in Ellen Frankel Paul, et al (1988), 173-221.

---. 1988b. "Moral Artifice." *Canadian Journal of Philosophy* 18:2, 385-418.

---. 1990. *Moral Dealing*. Ithaca: Cornell University Press.

---. 1990b. "Justice and Natural Endowment: Toward a Critique of Rawls' Ideological Framework." in *Moral Dealing* 150-170.

---. 1990c. "Bargaining and Justice." in *Moral Dealing* 187-206.

---. 1994. "Assure and Threaten." *Ethics* 104, 690-721.

Hammond, Peter. 1977. "Dynamic Restrictions on Metastatic Choice." *Economica* 44, 337-350.

Hampton, Jean. 1994. "The Failure of Expected-Utility Theory as a Theory of Reason." *Economics and Philosophy* 10:2, 195-242.

Hardin, Russell. 1982. *Collective Action*. Baltimore: Johns Hopkins University Press.

Harsanyi, John. 1965a. "Bargaining and Conflict Situations in the Light of A New Approach to Game Theory." *American Economic Review* 55:1, 447-457.

---. 1965b. "Rational-Choice Models of Political Behaviour vs. Functionalist and Conformist Theories." *World Politics* 21, 513-538.

---. 1976. *Essays on Ethics, Social Behaviour, and Scientific Explanation*. Dordrecht: D. Reidel Publishing.

---. 1977. *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations* New York: Cambridge University Press.

Hobbes, Thomas. 1651. *Leviathan*. Michael Oakeshott (editor). Oxford: Basil Blackwell, 1960.

Hume, David. 1739. *Treatise of Human Nature*. 1964 reprinting. L.A. Selby-Bigge (editor). Oxford: Clarendon Press, 1888.

Kavka, Gregory. 1983. "The Toxin Puzzle." *Analysis* 43:1, 33-36.

---. 1984. "Deterrent Intentions and Retaliatory Actions." in MacLean (1984), 155-159.

---. 1991. "Is individual choice less problematic than collective choice?" *Economics and Philosophy* 7:2, 143-165.

---. 1993a. "Rationality Triumphant." *Ethics*. 103, 349-351.

---. 1993b. "Internal Prisoner's Dilemma Vindicated." *Economics and Philosophy* 9:1, 171-174.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge, Mass: Harvard University Press.

---. "Devil's Bargains and the Real World." in MacLean (1984), 141-155.

MacLean, Douglas (editor). 1984. *The Security Gamble*. Totawa, NJ: Rowman & Allanheld.

McClennen, Edward. 1990. *Rationality and Dynamic Choice*. New York: Cambridge University Press.

Mele, Alfred. 1992. "Intentions, Reasons, and Beliefs: Morals of the Toxin Puzzle." *Philosophical Studies* 68, 171-194.

Morris, Christopher. 1988. "The Relation Between Self-Interest and Justice in Contractarian Ethics." in Ellen Frankel Paul, et al (1988).

Luce, R. Duncan, and Howard Raiffa. 1957. *Games and Decisions*. New York: Wiley.

March, James. 1978. "Bounded Rationality, Ambiguity, and the Engineering of Choice." *Bell Journal of Economics* 9:2, 587-607.

Moser, Paul. 1990. *Rationality in Action.* New York: Cambridge University Press.

von Neumann, John and Oscar Morgenstern. 1944. *The Theory of Games and Economic Behaviour*. Second Edition. Princeton, NJ: Princeton University Press.

Olson, Mancur, Jr. 1965. *The Logic of Collective Action.* Cambridge, Mass: Harvard University Press.

Derek Parfit. 1984a. *Reasons and Persons.* Oxford: Clarendon Press.

---. 1984b. "Rationality and Time." *Proceedings of the Aristotelean Society* 84, 47-82.

Ellen Frankel Paul, et al. (editors). 1988. *The New Social Contract.* New York: Basil Blackwell.

Rapoport, Anatol. 1966. *Two Person Game Theory.* Ann Arbor: University of Michigan Press.

Rawls, John. 1971. *A Theory of Justice.* Cambridge, Mass: Harvard University Press.

---. 1980. "Kantian Constructivism in Moral Theory." *Journal of Philosophy.* 77, 515-572.

Rorty, Richard. 1991a. *Objectivity, Relativism, and Truth.* Philosophical Papers Vol 1. New York: Cambridge University Press.

---. 1991b "The Priority of Democracy Over Philosophy." in *Objectivity, Relativism, and Truth* 175-196.

---. "Solidarity or Objectivity." in *Objectivity, Relativism, and Truth* 21-34.

Roth, Alvin. 1979. *Axiomatic Models of Bargaining.* Heidelberg: Springer-Verlag.

Savage, Leonard J. 1990. "Historical and critical comments on utility." in Moser (1990) 41-54.

Thomas Schelling. 1963. *The Strategy of Conflict.* Cambridge, Mass: Harvard University Press.

---. 1978. "Egonomics, or the art of self-management." *American Economic Review: Papers and Proceedings* 68, 290-294.

---. 1984a. "The Intimate Contest for Self-Command" in *Choice and Consequence* 57-82.

---. 1984b. *Choice and Consequence.* Cambridge, Mass: Harvard University Press.

Schick, Frederic. 1986. "Dutch Bookies and Money Pumps." *Journal of Philosophy* 83:2, 112-119.

Selten, Reinhard (editor). 1992. *Rational Interaction.* Berlin: Springer-Verlag.

Sen, Amartya. 1977. "Rational Fools: A Critique of the Behavioural Foundations of

Economic Theory." *Philosophy and Public Affairs* 6, 317-44.

Sidgwick, Henry. 1907. *The Methods of Ethics.* 7th edition London: MacMillan.

Simon, Herbert A. 1966. "Political Research: The Decision-Making Framework." in David Easton (1966) 15-24.

---, and James March. 1958. *Organizations*. New York: Wiley.

Strotz, Robert. 1955. "Myopia and Inconsistency in Dynamic Utility Maximization." *Review of Economic Studies* 23, 165-180.

Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation.* Cambridge, Mass: Harvard University Press.

Thaler, Richard. 1980. "Toward a Positive Theory of Consumer Choice." *Journal of Economic Behaviour and Organization* 1, 39-60.

Vallentyne, Peter (editor). 1991. *Contractarianism and Rational Choice*. New York: Cambridge University Press.