Social Science Information
http://ssi.sagepub.com/

Ulysses and the Sirens: A theory of imperfect rationality

Jon Elster

Social Science Information 1977 16: 469
DOI: 10.1177/053901847701600501

JON ELSTER

Ulysses and the Sirens: A theory of imperfect rationality *

> "... but you must bind me hard and fast, so that I cannot stir from the spot where you will stand me... and if I beg you to release me, you must tighten and add to my bonds" *(The Odyssey)*.

I

Ulysses was not fully rational, for a rational creature would not have to resort to this device; nor was he simply the passive and irrational vehicle for his changing wants and desires, for he was capable of achieving by indirect means the same end as a rational person could have realised in a direct manner. His predicament — being weak and knowing it — points to the need for a theory of *imperfect rationality* that has been all but neglected by philosophers and social scientists alike. Two path-breaking papers by R.H. Strotz[1] and George Ainslie[2] have laid the empirical and conceptual foundations on which all later work will have to build; in this paper I attempt a first step towards a synthesis. I also endeavour to broaden the empirical base itself, by adding some examples from fields not considered in these papers. Strotz mainly deals with examples from the theory of consumer behaviour, whereas Ainslie in his very wide-ranging discussion relies on findings from experimental psychology. Their work is summarized and discussed in Sections V and VII respectively. In Section VI the reader will find a discussion of a problem closely related to the one raised by Strotz. Sections III and IV adduce some examples from the history of philosophy, whereas Section VIII argues that many problems in political theory can be seen in this general perspective. The problem of binding oneself is also relevant for many questions in the philosophy of mind (who binds whom?) and in moral psychology (who has the right to bind whom?); these issues run through all Sections, with the possible exception of Section VIII.

II

It is not hard to think of everyday examples of persons binding themselves (or *precommitting* themselves, as I shall also say as a stylistic variation). In order to stop smoking it is standard practice to set up some causal machinery that will add force to your inner resolution: to tell your friends about your intention so as to invite their sarcastic comments if you are backsliding; to go for a walk in the mountains so as to make cigarettes physically unavailable; to cross the street when you see a tobacco shop further on so as not to be exposed to the sight of cigarettes; to take cold showers in order to strengthen your will power; to undergo hypnosis in order to induce aversion to tobacco; to make yourself believe that more cigarettes means certain death within five years. Similar but more complex problems are raised by obesity and alcoholism. Whereas the addiction to nicotine is a fairly isolated feature of the personality, being fat or alcoholic are character traits that are involved in most other traits as well; the desire to change them is a desire to become a different kind of person. Some further examples will be given. Persons with periodically recurring mental illness might bind themselves in advance, by issuing instructions that when the next episode occurs, the instructions that shall then be issued (refusal to be hospitalized or extravagant orders of fast cars) are not to be obeyed. Readers of *Lucien Leuwen* will recall the moment when Madame de Chasteller recognizes her love for Lucien and, fearing what it may lead her into doing, binds herself by choosing a companion who is certain not to permit her the smallest indiscretion. "Cet être si méchant me répondra de moi-même."[3] We may pause here to observe the subtleties that nearly always arise in these cases. Madame de Chasteller adduces a further motive for choosing a companion: the desire to punish herself for past indiscretions. "Et la sévérité de cette punition tranquillisa sa conscience : Madame de Chasteller se pardonna presque l'entrevue si légèrement accordée à Leuwen."[4] Self-punishment may indeed be a technique for binding oneself, through a self-induced process of operant learning that tends to eliminate the behaviour that is punished. In the present case, however, this would be redundant, because the punishment also has the effect of making that behaviour physically impossible. If Madame de Chasteller had punished herself by some other means, such as self-flagellation, one could impute to her a prudential motive, even if it would be hard not to suspect some neurotic component as well. In this case, however, prudence has led to the simpler strategy or reducing the number of feasible options, and then the motive for punishment can only be moral or neurotic.

I shall now propose some criteria for what it is to bind oneself. They are intended as necessary conditions only; an attempt to give necessary and sufficient conditions would be premature. An obvious requirement is that

(i) To bind oneself is to carry out a certain decision at time t_1 in order to increase the probability of another decision being carried out at time t_2.

The crucial point here is that the expected change in the probability of the later action must be the motive for the earlier one, not only a predictable and not unwelcome effect. If for example, Madame de Chasteller had chosen her companion in order to punish herself; if furthermore the binding effect of that choice was predictable and acceptable to her; and if finally she would not have chosen that disagreable person had she not committed her earlier indiscretions, even assuming that she could predict her later ones — then the first condition would not be satisfied and we would not have a case of binding oneself. If the psychological stance just described seems implausible, I can only answer that a similar attitude seems to underly the refusal of many persons to engage in complicated and strategic behaviour directed against themselves. For some persons, that is, the desire to stop smoking may be weaker than their desire not to be the kind of non-spontaneous person that could bring about a state of non-smoking through binding themselves. Such a person might engage in some activity knowing that it will lead to a state of non-smoking, on the condition that this is not the motivating purpose of that activity.

We may pause here to observe that this is *not* the problem that some (desired) results may be impossible to achieve except as by-products of activities undertaken for other ends; the person that sets out to obtain pleasure or to make himself into a cultured individual will usually be thwarted, unless the means become ends in themselves at some point during the process. Nor is it a question of cost-benefit analysis in the narrow sense that the necessary calculations and deliberations will consume more time and energy than will be saved if the strategy is successful. Rather the issue is one of cost-benefit calculus in a much broader sense: even if the deliberations do succeed in modifying the behaviour or the character in the desired way, the very activity of deliberating can modify the character for the worse and in ways judged even more important, through the stultifying effects on spontaneity. This calculus would in itself be a "planning of spontaneity" that might seem rather paradoxical, but I agree with Thomas Nagel [5] that if spontaneity is a *value* then it is absurd not to protect it.

Condition *(i)* permits more complex and hierarchical methods, involving three or more decisions. If, for example, the usual pattern of manic-depressive episodes were depression succeeded by exaltation (in fact the pattern is the opposite one), then the following case might be envisaged. At time t_1 the person, being in the "normal" or "ground" state, issues the instructions that the instructions he will issue at time t_2, during the depressive phase, to the effect that the instructions he will issue at time t_3, during the manic phase, are not to be obeyed, are not to be obeyed. This assumes that his considered opinion is that a person in a manic phase is not necessarily irresponsible, or that treating him as irresponsible will aggravate the long-term problem even if easing the short-term one. In somewhat fanciful terms we might speak here of an alliance between the early and the late self against the intermediate and more docile self. (Three-step techniques are also explored by Ainslie [6], but in his examples the point is that the early and late stage may combine against the middle one so as to make precommitment superfluous.) In this case the "ground state" can be clearly distinguished from the pathological ones, so that external observers would know which instructions to follow, but this might not always be so. If a person one day issues an instruction that future instructions of some specific kind are not to be obeyed, and then retracts these meta-instructions the following day, what would be the criteria for deciding whether the self making the retractation is a more authentic self than the one that issued the meta-instructions, or the very self whose lack of authenticity made those meta-instructions necessary?

According to the first criterion, taken by itself, any act of investment, *i.e.* any sacrifice of present goods in order to make more goods available later on, would count as binding oneself. This, I think, is counter-intuitive. It might be necessary to bind oneself in order to make that sacrifice, *i.e.* to make a decision at t_1 in order to increase the probability of the sacrifice at t_2, but the sacrifice at t_2 is not an act of binding oneself even though it increases the probability of certain consumption decisions being carried out at t_3. In order to exclude this and similar cases, I shall impose the further requirement that:

> *(ii)* If the act at the earlier time induces a change in the set of options that will be available at the later time, this counts as binding oneself only if the change is a *reduction* of the feasible set.

In Section IX below I argue that inducing changes in the feasible set is but one of many strategies for binding oneself; the second requirement states that only contractions of the set shall count as instances of *this* strategy. Once again we must insist upon the distinction between motives and predictable desired effects. An act of investment may be irreversible (though it need not be), and as such may have the consequence of binding the investor to some particular allocation of consumption over time (Section V below). This may even be part of the motive for the investment (so that the first criterion is satisfied), but it need not be.

A further criterion is that

> *(iii)* The effect of carrying out the earlier decision must be to set up some causal process in the external (material) world.

This excludes, for example, decisions to decide. The point here is not that the carrying out of a decision to decide simply is to carry out the later decision itself; I feel that a decision to decide can be carried out by setting up some mental attitude that can be described as "resolution", "firmness of purpose", etc. Nor is the rationale behind criterion *(iii)* that decisions to decide are without efficacy, though I think that they in general have very little impact. "I decide that I shall decide that p" has the same ritual and redundant sound as "if someone were to buy several copies of the morning paper to assure himself that what it said was true" [7]. Rather the point is that our intuitive notions of what it is to bind oneself seem to require that we temporarily deposit our will in some external structure; that we set up a causal process in the external world that after some time returns to its source and modifies our behaviour. Criterion *(iii)* also has the consequence of excluding what Ainslie calls "private side bets" [8], which essentially is a decision to group future rewards so that they stand or fall together. Whatever the empirical importance or the psychological plausibility of this ingenious strategy, the fact that it is a response to the same problem that gives rise to the technique of binding oneself, should not lead us into saying that it is an instance of that technique. Private side bets, like the "strategy of consistent planning" that will be considered in Section V, is an alternative to the method of precommitment.

A fourth and very important requirement is that:

> *(iv)* The resistance against carrying out the decision at t_1 must be smaller than the resistance that would have opposed the carrying out of the decision at t_2 had the decision at t_1 not intervened.

To take cold showers in order to develop the strength of will that shall enable me to stop smoking, is not a very good strategy if stepping under the cold shower requires the very willpower it is supposed to develop. The problem underlying condition *(iv)* is the same as the general problem to which the strategy of Ulysses is a solution: our general resistance to uphill walking, and our preference for downhill strolls. I have argued elsewhere [9] that the capacity for *waiting* and for using *indirect strategies* ("one step backwards, two steps forwards") is among the features that distinguish man from the other animals, but this is not to say that the capacity is always fully utilized. Only man is in the possession of a general capacity for seeking global maxima at the expense of local maxima; but it is equally true that only man is susceptible of *akrasia*, because that notion only makes sense on a background of successful long-term planning and successful resistance to temptation in many cases; a fortiori only man is capable of overcoming his weakness of will through the use of techniques such as precommitment, private side bets, consistent planning, etc.

This implies that precommitment in some cases can be seen as an indirect strategy that enables us to use indirect strategies. If I am unable to take a whole step backwards, then I might perhaps be able to take half a step, knowing that after that first half step I shall be committed to the second half and enabled to take the two steps forward later on. As already stated, investment is a case in point. If I have designed a new type of fishing net that will enable me to catch twice as much as the old one, and if I know that while I make the new net I shall get so hungry that I shall prefer to go fishing with the old and as a consequence never get enough time for uninterrupted work on the new, then *destruction of the old net* could be a rational choice: an indirect strategy compelling me to use the indirect strategy of investing in a new net. This is not simply a case of burning one's bridges, for if one intends to go forward these bridges serve no rational purpose anyway. Destroying useful means of production that could have eased the strain during the construction of the new, is a more radical procedure. It is sometimes suggested that the rapid economic growth in societies that have been ravaged by war, such as Japan or Germany, should be seen in this perspective. I am more than sceptical about this notion. It may be true that the destruction of outdated machinery has the *effect* of enabling a nation to escape "the penalties for taking the lead", but on all interpretations of history but the conspiratorial one this would never be a *motive* for undertaking a war, as it would have to be for the first criterion above to be satisfied. (In Section VIII below I shall have something more to say about the even more obscure suggestion that such precommitment could be the *function* of war.)

The general problem is the following: for a given task of uphill walking (which I resist with a given force) and for a given strength of will, does there always exist a strategy of precommitment that *(i)* is within my will power and *(ii)* is capable of getting me ultimately to the top of the hill? ("Ultimately", because we might envisage a sequential hierarchy of acts of binding oneself, each of which requires a relatively small effort.) It has been argued (notably by Descartes, as we shall see in Section IV) that anyone can achieve anything if he goes about it in a sufficiently roundabout manner. In a special case and in a special sense (to be discussed in Section VI) this holds demonstrably even for persons with zero willpower. If we accept, at least for the sake of argument, that we are all-powerful in this sense, how should we then explain the fact that we do not always use these options that are available to us? By hypothesis we cannot invoke weakness of will. We can, however, appeal to cost-benefit considerations, in either the narrow or the broad sense distinguished above. For one individual the very act of deliberating might require an effort that lies above the threshold of his willpower; it does not then make any difference whether he is capable of the effort required to carry out this comparison itself. (The situation is somewhat analogous to Maxwell's demon, who could surmount the second law of thermodynamics only on the assumption that calculation is costless and frictionless.) For another individual the process of getting to the top (which he wants) would also induce other character changes (which he resists), in which case it could be rational to refuse the option that is assumed to exist.

The last condition I shall discuss is:

> *(v)* The act of binding oneself must be an act of commission and not of omission.

In order to see the need for some such condition, consider the following problem in the philosophy of education. I suspect that a not uncommon vision of the development of moral and intellectual faculties in the child is something like a four-stage sequence. In a first stage the power of parent and teacher is just a brute fact accepted only because there is no other alternative. In a second stage the child is still powerless to reject authority, but rational enough to see that he would not have wanted to do so even had he been able to. In a third stage the child is (physically and legally) capable of rejecting authority, but refrains from doing so because he sees that it is in his own interest to remain bound. In a final stage liberation takes place at the common initiative of all parties. Now whatever the moral appeal of this fairy tale, I would not like to count the third (and a fortiori not the second) stage as an example of binding oneself. The fact that someone prefers not to leave a given state is not evidence that he would freely have entered that state; there are transaction costs and uncertainties involved that destroy the apparent symmetry of exit and entry. As we shall see in Section VI below, preferences are always relative to a past history of choices; if the child had known (from experience) the states to which he prefers the state of being bound, his preferences might have been very different. I shall return to the implications of criterion *(v)* in the discussion of abdication in Sections IV and VIII below.

III

Pascal's wager has an obvious relevance for our problem. The argument has two parts, of which the first and most well-known goes roughly like this: since there is a certain positive probability that God exists, and since he that believes in the existence of God receives an infinitely large gain if he proves right, whereas only a finite amount is to be staked, the principle of expected utility maximization requires that one should believe. (This, by the way, assumes that eternal bliss is not reduced to present value at some positive discount rate, as it is in a recent extravaganza.)[10] Now "belief" is a very particular kind of action, in that it cannot be performed just on the will's saying so. Whence the necessity of the second step in the argument:

> "Vous voulez aller à la foi, et vous n'en savez pas le chemin; vous voulez vous guérir de l'infidélité, et vous en demandez le remède: apprenez de ceux qui ont été liés comme vous, et qui parient maintenant tout leur bien; ce sont gens qui savent ce chemin que vous voudriez suivre, et guéris d'un mal dont vous voulez guérir. Suivez la manière par où ils ont commencé : c'est en faisant tout comme s'ils croyaient, en prenant de l'eau bénite, en faisant dire des messes, etc. Naturellement même cela vous fera croire et vous abêtira. — 'Mais c'est ce que je crains.' — Et pourquoi? Qu'avez-vous à perdre?"[11]

Initially there is no reason for believing, only a reason for making yourself believe. The causal efficacy of a belief for a given end can never provide grounds or reasons for *adopting* that belief, with the possible exception of self-fulfilling beliefs[12]. Nevertheless this efficacy might give a reason for *precommitting* oneself to the belief, in the sense of setting up a series of actions that will have the predictable result of my coming to believe. In Pascal's argument these actions can be described as "going through the motions"; acting as if one believed in order to generate the real thing.

Pascal's self-directed argument may be contrasted with the other-directed argument that comes more naturally to the social scientists. Arthur Stinchcombe offers the following perceptive analysis:

> There are two possible causal links between control over activities in the present and the structure of activities and values in the future. *(a)* Activities established by current power-holders, embodying their values, may serve other functions than serving those values. Such additional functions will preserve the activities, even if commitment to the value is low. *(b)* People become committed to what they are doing, perhaps in order to reduce cognitive dissonance, so that one way to socialize people is to get them to act in terms of that value without belief and allow belief to follow[13].

The second of the two causal links is also at the heart of Pascal's argument with the difference that in his case the individuals themselves are being asked to manipulate their own beliefs with action as an intermediate link. That rulers or ecclesiastics can exploit the pomp of religious ceremonies to bolster religious faith against the attack of reason, is of course well known. "Il y eut un *Te Deum*, des flots d'encens, des décharges infinies de mousqueterie et d'artillerie; les paysans étaient ivres de bonheur et de piété. Une telle journée défait l'ouvrage de cent numéros des journaux jacobins."[14] It is vastly more paradoxical that reason itself should adopt these methods against itself. Let us take a closer look at some aspects of this paradox.

Bernard Williams has argued convincingly that even if it is possible to decide to believe p, one cannot both believe p and believe that the belief that p stems from a decision to believe p[15]. If the decision to believe p is to be carried out successfully, it must also obliterate itself from the memory of the believer. The point is not that the belief that p is incompatible with the belief that the belief that p is the result of some causal process; all beliefs are causally produced. Rather the idea is that in the case of a decision to believe, the belief stems from the wrong sort of causal process, *i.e.* from a process quite unrelated to any grounds or reasons for believing. The implication of this argument is that the decision to believe can only be carried out successfully if accompanied by a decision to forget. This, of course, is just as paradoxical as the decision to believe:

> The Heart cannot forget
> Unless it contemplate
> What it declines[16].

You can *make* someone forget something, but not by telling him to forget it, which quite probably will have the opposite result, particularly if the injunction is taken seriously[17]. A fortiori you cannot simply decide to forget something either; unless, once again, you bind yourself in some manner by setting up some roundabout machinery to induce forgetfulness. The most efficient procedure would clearly be to start up a single causal process with the double effect of inducing belief *and* making you forget that it was ever started up. Asking to be hypnotized is one such mechanism; according to Pascal acting as if one believes is another, for "cela vous fera croire *et* vous abêtira". The loss of the critical faculty is not simply a *by-product* of the self-induced faith; it is an essential *condition* for that faith to be held seriously, and if it had not followed from the faith-inducing process itself a separate process would have been needed to bring it about.

Now even if this may lead to the faith being held seriously, one wonders whether it should be taken seriously. It would have to be a very tolerant or uncritical God that bestowed the grace upon someone whose faith had been obtained in this manner; it would be rather like praising someone for being right while knowing that he was right for the wrong reasons. Now *after the event* (*i.e.* after the onset of faith) someone might say that the religious practice of acting as if one believed only had the effect of triggering off a faith which in retrospect was latently present all the time, and he might claim that the genesis of the faith and the fact that it was acquired for the wrong reasons, should be irrelevant if in fact it is held for the right reasons. I submit, however, that no one could accept this claim *before the event*, *i.e.* at the time when the wager has to be accepted or rejected. For either this claim would only be a piece of self-deception which could not deceive God, or it would be an authentic one that could only be accepted by someone who had already acquired an authentic faith. It may be possible to pull oneself up by the bootstraps, but no one can rationally count on being able to do so.

Unless I am mistaken there is also, however, a second line of argument that is hard to square with the wager. In the wager it is assumed that reason in itself cannot offer any argument for the existence of God, only for the utility of believing in that existence. In other contexts Pascal seems to think that reason is capable of *arriving at* that belief by ratiocination, but that it needs the help of the passions in order to *maintain* belief:

> "Car il ne faut pas se méconnaître : nous sommes automates autant qu'esprit; et de là vient que l'instrument par lequel la persuasion se fait n'est pas la seule démonstration. Combien y a-t-il peu de choses démontrées! Les preuves ne convainquent que l'esprit. La coutume fait nos preuves les plus fortes et les plus crues; elle incline l'automate, qui entraîne l'esprit sans qu'il y pense [...] Enfin il faut avoir recours à [la coutume] quand une fois l'esprit a vu où est la vérité, afin de nous abreuver et nous teindre de cette créance, qui nous échappe à toute heure; car d'en avoir toujours les preuves présentes, c'est trop d'affaire. Il faut acquérir une créance plus facile, qui est celle de l'habitude, qui, sans violence, sans art, sans argument, nous fait croire les choses, et incline toutes nos puissances à cette croyance, en sorte que notre âme y tombe naturellement. Quand on ne croit que par la force de la conviction, et que l'automate est incliné à croire le contraire, ce n'est pas assez. Il faut donc faire croire nos deux pièces : l'esprit, par les raisons,

> qu'il suffit d'avoir vues une fois en sa vie; et l'automate par la coutume, et en ne lui permettant pas de s'incliner au contraire." [18]

This seems essentially to be the classical Aristotelian view of moral education: you become a good man by performing good actions. Initially the performing of these actions is an uphill climb which requires efforts and meets with resistance, but the result is to set up a *hexis* such that the same actions flow as naturally from it as water seeking its downhill course [19]. The distinction between knowing something in the abstract and knowing it in the sense of having the proof before your mind, is also used by Aristotle in his solution to the problem of *akrasia* [20]. The *akrates* knows what he should do, but only in the sense in which a sleeping geometer can be said to "know" a geometrical theorem. One cannot constantly *keep* before one's mind all that one knows, even if one is able to *bring* it to mind under the appropriate conditions, which may be lacking in the heat of action. Now this is a "solution" to the problem of weakness of will only in the sense of explaining how that phenomenon is at all possible [21]; it is not a solution in the sense of a recipe for overcoming it. A solution in the latter sense is forthcoming if we bring the Aristotelian theory of moral education to bear upon the problem of *akrasia*; we then get the view exposed by Pascal in the test quoted above. According to this view, reason has a double task. In the first place only reason can arrive at the intellectual insight that God exists. In the second place reason knows that intellectual insight is not enough, and therefore starts up a series of actions that will engender "automatic" or "customary" belief, permitting me to act rightly without having to mobilize at each instant the whole battery of arguments.

It should be fairly clear that this argument is very different from the wager argument. Let us try to bring out the difference in a more explicit manner. Writing belief$_r$ for the rational belief arrived at by ratiocination, belief$_c$ for customary belief and "p" for the proposition "God exists", the wager argument goes like this:

1. I do not believe$_r$ that p
2. I believe$_r$ that I should always believe$_c$ that p or believe$_r$ that p
3. I do not believe$_r$ that I shall come to believe$_r$ that p
4. I believe$_r$ that it is impossible (to believe$_c$ that p and to believe$_r$ that I believe$_c$ that p because of a decision to believe$_c$ that p)
5. Therefore I decide to act so as to bring it about (*i*) that I believe$_c$ that p and (*ii*) that I do not believe$_r$ that I believe$_c$ that p because of a decision to believe$_c$ that p.

Here premise (1) is the very starting point for the argument, which addresses itself to an unbeliever. Premise (2) is the conclusion of what I have called the first part of the wager argument, *i.e.* the reasoning in terms of expected utility maximization. Premise (3) follows, as argued above, from Premise (1). Premise (4) is given by Williams' argument. Premises (2) and (3) give part (*i*) of the conclusion (5); part (*i*) and Premise (4) then give part (*ii*). By contrast the Aristotelian argument goes like this:

6. I believe$_r$ that p
7. I believe$_r$ that I should always believe$_c$ that p or believe$_r$ that p
8. I believe$_r$ that it is impossible always to believe$_r$ that p
9. Therefore I decide to act so as to bring it about that I always believe$_c$ that p.

From the crucial difference between premises (1) and (6) stem several further differences between the two arguments. Premises (2) and (7) are identical in wording, but the word "should" does not reflect the same kind of rationality in the two cases. (6) is in itself a sufficient reason for (7), whereas (2) can only be sustained by invoking the causal efficacy of the belief. Premises (3) and (4) correspond in some sense to premise (8), but the logic is very different. That (3) follows from (1) is conceptually necessary, as is also the premise (4); by contrast, premise (8) only expresses a psychological difficulty. Both arguments conclude that one should bind oneself to a belief$_c$, but the wager argument also has the added implication that one should bind oneself to forget the argument itself.

There is a phrase in *Les Pensées* that could be used as a tag for both arguments: "Il n'y a rien de si conforme à la raison que ce désaveu de la raison." [22] It is clear by now that this disavowal can either be taken in the strong sense of an *abdication of reason* or in the weaker sense of the *insufficiency of reason* alone. In the next Section we shall see that Descartes' critique of reason probably is weaker than either of these; here we should insist on the fact that even the abdication of reason is a weaker critique than the *dethronement of reason* that we associate with the names of Nietzsche and Freud. In the wager argument *it is reason itself that decides to abdicate;* a second-level rationality deciding that rationality should be abolished. The dethronement of reason is much more radical, because it assumes that all forms of rationality really are nothing but dressings-up of vital drives; in fact it is misleading even to speak of dethronement, because reason never had governed in the first place. The dethronement of reason is only to effectuate *de jure* what always has been the case *de facto:* the supremacy of the passions.

IV

To the phrase by Pascal quoted at the beginning of the last paragraph corresponds this statement by Descartes: "la principale finesse est de ne vouloir point du tout user de finesse" [23]. In this Section the main task will be to propose an interpretation of this phrase and, more generally, of the Cartesian theory of rational choice. As already stated, we shall see that the critique of reason ("finesse") implied by that statement is a very weak one; it can be briefly characterized as *a critique of instant rationality.* I shall also argue that even though this critique of instant rationality does not imply the insufficiency of reason, it is not implausible to see a psychological connection between the two. For this purpose I shall start out by showing that Descartes does in fact have a theory, set forth in *Les passions de l'âme*, which is very similar to the Aristotelian argument adduced by Pascal. Even though Descartes himself never makes that theory bear upon the theory of instant rationality, the fact that he holds both of these views and that there is a plausible manner of connecting them would seem to justify the interpretation proposed towards the end of this Section.

Towards the end of the first part of *Les passions de l'âme* Descartes draws distinction between two ways of fighting the passions:

> "... ceux en qui naturellement la volonté peut le plus aisément vaincre les passions et arrêter les mouvements du corps qui les accompagnent ont sans doute les âmes les plus fortes; mais il y en a qui ne peuvent éprouver leur force parce qu'ils ne font jamais combattre leur volonté avec ses propres armes, mais seulement avec celles que lui fournissent quelques passions pour résister à quelques autres. Ce que je nomme ses propres armes sont des jugements fermes et déterminés touchant la connaissance du bien et du mal, suivant lesquels elle a résolu de conduire les actions de la vie." [24]

The will may overcome the passions through sheer will power, but it may also use the indirect strategy of pitting the passions against each other. Albert Hirschman has recently explored [25] the widespread use of this notion in the seventeenth century, but he does not, perhaps, distinguish with sufficient clarity between the analytical, the manipulative and the strategic purposes to which it may be harnessed. The analytical purpose would be to determine to which extent the passions actually tend to neutralize each other *in men as they are.* The manipulative and the strategic approaches would try to set passion against passion in order to modify the behaviour: the behaviour of others in the manipulative case, the behaviour of self in the strategic case. This modification could operate in either of two ways. In the first place person A might try to influence person B (leaving open that A and B may be the same person) by associating some additional reward or punishment to some of the courses of action available to B. For the case that A = B, this is the method of binding oneself through public side bets [26]. In the second place, A might want to change the character or the system of passions of B, so that even with a constant reward system a different choice would be made. Or to put the matter differently: to pit passion against passion may take the form of changing the situation so as to bring new passions to bear upon it, or the form of changing the passions so as to bear upon the situation in a new manner. I believe that all the cases adduced by Hirschman, in addition to the purely analytical ones, come in the category of reward-manipulation. I also believe that Descartes' idea is that behaviour modification can take place through strategic (self-directed) character modification. The following passage should sustain this interpretation:

> "Et il est utile de savoir que, comme il a déjà été dit ci-dessus, encore que chaque mouvement de la glande semble avoir été joint par la nature à chacune de nos pensées dès le commencement de notre vie, on les peut toutefois joindre à d'autres par habitude [...] Or ces choses sont utiles à savoir pour donner le courage à un chacun d'étudier à regarder ses passions; car, puisqu'on peut, avec un peu d'industrie, changer les mouvements du cerveau dans les animaux dépourvus de raison, il est évident qu'on le peut encore mieux dans les hommes, et que *ceux même qui ont les plus faibles âmes pourraient acquérir un empire très absolu sur toutes leurs passions*[27]".

Anyone can do anything; the smallest amounts of will-power suffice for the most extraordinary feats of self-control, given an understanding of the physiological mechanisms by which habits are formed and changed. *Hexis* in Aristotle, custom in Pascal and habit in Descartes are all seen as the end result of non-habitual actions. The result may be an unintended and unforeseen one, if one is ignorant of the causal process by which it is produced; if one has insight into the workings of the mind (as according to Aristotle one should have)[28] it can be exploited for the purpose of self-education and self-control.

I believe to have shown that the theme of precommitting oneself is indeed present in Cartesian psychology, but it is certainly less important than another theme which I shall now proceed to discuss. This theme — the critique of instant rationality — is first found in a famous passage from the *Discours de la Méthode:*

> "Ma seconde maxime était d'être le plus ferme et le plus résolu en mes actions que je pourrais, et de ne suivre pas moins constamment les opinions les plus douteuses, lorsque je m'y serais une fois déterminé, que si elles eussent été très assurées. Imitant en ceci les voyageurs qui, se trouvant égarés en quelque forêt, ne doivent pas errer en tournoyant tantôt d'un côté, tantôt d'un autre, ni encore moins s'arrêter en une place, mais marcher toujours le plus droit qu'ils peuvent vers un même côté, et ne le changer point pour de faibles raisons, encore que ce n'ait peut-être été au commencement que le hasard seul qui les ait déterminés à le choisir; car, par ce moyen, s'ils ne vont justement où ils désirent, ils arriveront au moins à la fin quelque part où vraisemblablement ils seront mieux que dans le milieu d'une forêt"[29].

I shall follow a two-step procedure in interpreting this text. In the first step I sketch a number of possible readings, two of which are singled out as equally plausible and as equally different from Pascal's critique of reason. In the second step I adduce some further passages from Descartes' writings that tend to attenuate the contrast between Descartes and Pascal, even though I believe that a genuine difference remains. Throughout the discussion I draw upon modern economics and game theory in order to bring out the very real implications of what might otherwise have seemed to be abstract and scholastic distinctions.

Now *the first interpretation* of the passage that comes to the mind of an economics-minded reader is, I think, the following: a continuous evaluation and reevaluation of the situation requires so much time that it can be expected to more than outweigh the time gained by the improved direction that issues from the evaluation. The traveller will get more quickly out of the forest if he follows some — perhaps arbitrarily chosen — straight line than if he constantly halts in order to adjust his direction. This interpretation is very close to the attempts that have been made in recent years [30] to reduce Herbert Simon's theory of "satisficing" or of "limited rationality" [31] to a species of maximizing or total rationality. This reduction stresses that the use of rules of thumb or stereotyped decision principles, such as "always follow a straight line when lost in a forest", can be optimal if they permit us to economize on the cost of gathering and evaluating information. To seek the abstractly optimal solution to a technical problem, for example, might require so much time and money that the firm would go bankrupt long before the solution emerged. The reason why French firms regularly lose contracts to American firms have been interpreted along these lines; the excessive rationality of the French becomes an obstacle to economic rationality. I have argued elsewhere [32] that Leibniz's rationalism, or rather his attempt to translate philosophical rationalism into economic rationality, also is profoundly irrational from the practical point of view; to the extent that the first interpretation of Descartes is correct, one could say that his moral philosophy is an anticipated answer to Leibniz.

I do indeed believe that this interpretation is one of the two that with roughly equal plausibility may be imputed to Descartes. Before I go on to the other interpretations, I would like to stress that the decision to follow a straight line is a genuine *decision;* an act of commission and not of omission. It is made before the event, and is not just a rationalization made up (or imputed to the actor) after the event and having the occurrence of the event as its sole or main evidence. This point is important in view of the numerous pseudo-explanations that have been offered in recent years in order to prove that all kinds of apparently irrational behaviour "really" are governed by the principle of rational choice. Douglass North has argued, for example, that *ideologies* "are a way of economizing on the costs on information and therefore are in general a rational response" [33] because they spare us the trouble of evaluating each situation separately and on its merits. Now someone might possibly adopt an ideology, such as a rigidly negative attitude towards a minority group, for such reasons, but if this kind of explanation shall have force one would have to locate the decision in space and time. To say that the fact of *not jettisoning* an ideology that one has taken over from one's parents is evidence that one has *deliberately adopted* the ideology in order to economize on the costs on information, would be an instance of the same fallacy that we discussed in the last paragraph of Section II above. One should also be extremely on guard against the notion that an ideology could be rational even though not deliberately adopted, in the sense of being a *functional* response to a given problem. As I have argued in more detail elsewhere [34], only some kind of natural selection of ideologies could bring about this non-intentional adaptation; and even if such selection is not unthinkable in societies, it is most implausible in the present case.

In the preceding paragraphs we have distinguished between four approaches to rational choice: abstract or maximizing rationality (Leibniz), limited or satisficing rationality (Simon), "satisficing-as-maximizing" (Descartes: first interpretation) and "maximizing-as-satisficing" (natural selection theories of rationality). In a very important and strangely neglected article S.N. Winter has offered some very powerful arguments that tend to demolish all three alternatives to Simon's approach; these arguments also suggest *the second interpretation* of Descartes. Winter observes that the attempt to reduce satisficing to maximizing gives rise to an infinite regress, because "this choice of a profit maximizing information structure itself requires information, and it is not apparent how the aspiring profit maximizer acquires this information, or what guarantees that he does not pay an excessive price for it" [35]. Take the case of a multinational firm that decides not to enter the forward exchange market because the information costs of the operation would exceed the benefits [36]. Then we shall have to ask how the firm decided how much information to acquire before taking the decision not to acquire the information needed for the forward exchange operation. Unless one could prove (and I do not see how one could prove) that the deviation from the "real" optimum converges to zero or at any rate rapidly becomes smaller for each new level in the information hierarchy, this argument not only has the implication that in every decision there must be a cut-off point where calculation stops and you simply have to make an unsupported choice, but also that this point might just as well be as close to the action itself as possible. Why, indeed, seek for precision in the second decimal if you are uncertain as to the first? This line of argument could support a *decisionist* interpretation of Descartes which would indeed imply a rather radical critique of reason. The only task of reason would then be to prove the impossibility theorem just sketched, and from that point onwards intuition, decision or satisficing would be on their own, unsupported by formal reasoning.

I believe that this interpretation is the least plausible of the three readings I am offering of Descartes's moral philosophy, because it is hard to fit in with the texts (to be quoted below) dealing with the calculus of self-interest. Nevertheless there seems to be a strand in Descartes' writings, and indeed in his character, that reflects this preference for the unsupported decision that is rigidly adhered to once taken. Even in the absence of the impossibility theorem, this preference could be supported by an aristocratic distaste for calculations of any kind and by an equally aristocratic predilection for absolute firmness of character, however eccentric. It is hard to read the correspondence of Descartes without being struck by this "grand seigneur" aspect of his nature: never explain, never apologize.

I now pass to the *third interpretation*. This would be to argue that not only the *time* spent getting out of the forest is increased if the traveller constantly stops to reevaluate the situation, but that this instant rationality actually makes the *path* itself longer than it would have been along an arbitrarily (or at least along an optimally) chosen straight line. In order to bring out the mechanism involved here, we shall see how various economists have discussed a very similar problem in the context of finding the optimal use of forecasts in planning. Following Milton Friedman, Clem Tisdell has argued, in a language very similar to the one we use here, that:

> "... even if the policy-maker has some (positive) ability to predict, it may nevertheless be optimal for him to follow an inflexible policy. Indeed, [this paper] shows that, *ignoring the increased decision costs* which 'fine-tuning' policies may involve and the unfavourable uncertainty effects which they may generate in a group, deficiencies in the ability to predict may, even if there is (considerable) ability to predict, make flexible or zig-zag policies undesirable [...] Patterns emerge which are helpful from a prescriptive point of view and which also explain the rationality of actual behaviour that might otherwise appear to be irrational, *e.g.* the decision-maker who has predictive ability but does not adjust to his short-term predictions may well be acting rationally" [37].

In the phrase I have italicized Tisdell explicitly distinguishes his own approach (corresponding to the third interpretation of Descartes) from the cost-of-decision approach (corresponding to the first interpretation). In a later article Leif Johansen [38] has shown that matters are not quite as simple as Descartes-Tisdell assume them to be. The choice is not between following an initial decision as if it were absolutely correct and acting upon successive forecasts as if each of them were absolutely reliable; rather the optimum would be to steer a middle course and take some account of the changing forecasts while not changing direction to the extent that one would have done if the current forecast was always thought to give the correct value. This technical point apart, Tisdell's argument seems conceptually sound.

One of the examples offered by Tisdell is the inflation-unemployment trade-off (the Phillips curve). This problem is also discussed by William Nordhaus, in his important work on the political business cycle. I shall argue that Nordhaus' approach is closer to the problem of binding oneself, which by contrast is not very relevant to the Descartes-Tisdell argument. Tisdell shows that under certain conditions the optimal policy may be to hold a constant level of unemployment rather than to adjust it to current forecasts. Nordhaus' problem is quite different: his policy makers do not have to abstract from current forecasts, but from electoral considerations. The main thrust of his analysis seems to be this: voters and planners always attach some weight to the welfare of later generations, whereas politicians — trying to maximize votes at the next election — are exclusively concerned with the present. This has two consequences: in the long run society will choose a policy with lower unemployment and more inflation than is optimal; in the short run each electoral period will begin in austerity and end in potlatch. Among the possible remedies to this bias in the democratic system Nordhaus discusses the strategy of Ulysses:

> "A third possibility is to entrust economic policy to persons that will not be tempted by the Sirens of partisan politics. This procedure is typical for monetary policy, which for historical reasons is lodged in the central banks (as in the independent Federal Reserve System in the US or the Bank of England). A similar possibility is to turn fiscal policy over to a Treasury dominated by civil servants. It may be objected, however, that delegating responsibility to an agency that is not politically responsive to legitimate needs is even more dangerous than a few cycles. This danger is frequently alleged regarding central banks which pay more attention to the 'soundness of the dollar' or the latest monetarist craze than to fundamental policy problems .[39]"

I shall return to this strategy in Section VIII below. Here I shall only observe that it corresponds to a paraphrase of Pascal's formula: "Il n'y a rien de si conforme à la politique que ce désaveu de la politique." To remove monetary policy from the political sphere would itself be a political act; the abdication of politicians and not their overthrow. In his work on the same problem Assar Lindbeck also suggests depolitization of some policy agencies as a possible remedy; among the other possible strategies listed I was especially struck by his suggestion that elections could be randomly spaced so that politicians would not have an incentive for turning policies into means for electoral purposes[40]. Albert Hirschman remarks that "unpredictability is power"[41]: you can control your environment by making yourself appear unpredictable. Conversely you may control yourself by making the environment an unpredictable one. (*The Lottery* by Borges pursues this notion to its logical conclusion and beyond.)

As already stated, I do not think it possible to say whether the first or the third interpretation of Descartes is the most plausible one. They have in common the critique of instant rationality which I take to be the central point he is trying to make. There is virtually no question of a *temptation* to be resisted; the problem is ignorance rather than weakness of will. The ignorance of someone who has never learnt geometry is different from the ignorance of the sleeping geometer; once we have acquired the insight that it may be rational not to adhere to the instant rationality, there is nothing that prevents us from following that insight. Such is, at any rate, the preliminary conclusion that follows from a reading of the cited text from *Discours de la Méthode* taken by itself. I shall now adduce some further passages that may suggest a slightly different conclusion.

In Descartes' writings there is an interesting and (as far as I know) quite neglected analogy between his cosmological views and his views on society. Common to both are the tenets of *atomism* and of *optimism*: erratically moving particles will by themselves create a perfectly ordered universe; egoistically motivated individuals will act so as to further the common good. I shall quote two passages so similar in wording that it is hard not to think that they have their source in some common inspiration:

> "*Car Dieu a si merveilleusement établi ces Lois*, qu'encore que nous supposions qu'il ne crée rien de plus que ce que j'ai dit, et même qu'il ne mette en ceci aucun ordre ni proportion, mais qu'il en compose un Chaos, le plus confus et le plus embrouillé que les Poètes puissent décrire: elles sont suffisantes pour faire que les parties de ce Chaos se démêlent d'elles-mêmes, et se disposent en si bon ordre qu'elles auront la forme d'un Monde très parfait." [42]

> "J'avoue qu'il est difficile de mesurer exactement jusqu'où la raison nous ordonne que nous nous intéressions pour le public; mais aussi n'est-ce pas une chose en quoi il soit nécessaire d'être fort exact: il suffit de satisfaire à sa conscience, et on peut en cela donner beaucoup à son inclination. *Car Dieu a tellement établi l'ordre des choses*, et conjoint les hommes ensemble d'une si étroite société, qu'encore que chacun rapportât tout à soi-même, et n'eût aucune charité pour les autres, il ne laisserait pas de s'employer ordinairement pour eux en tout ce qui serait de son pouvoir." [43]

The two phrases that I have italicized must proceed from the same mould of thought; the underlying logic as well as the actual choice of words are too similar for any other conclusion to be possible. Let us now see how Descartes elaborates upon the sociological theory implied by the second of the above passages:

> "La raison qui me fait croire que ceux qui ne font rien que pour leur utilité particulière, doivent aussi bien que les autres travailler pour autrui, et tâcher de faire plaisir à un chacun, autant qu'il en est en leur pouvoir, s'ils veulent user de prudence, est qu'on voit ordinairement arriver que ceux qui sont estimés officieux et prompts à faire plaisir, reçoivent aussi quantité de bons offices des autres, même de ceux qu'ils n'ont jamais obligés, lesquels ils ne recevraient pas, si on les croyait d'autre humeur, et que les peines qu'ils ont à faire plaisir, ne sont point si grandes que les commodités que leur donne l'amitié de ceux qui les connaissent. Car on n'attend de nous que les offices que nous pouvons rendre commodément, et nous n'en attendons pas davantage des autres; mais il arrive souvent que ce qui leur coûte peu, nous profite beaucoup, et même peut nous importer de la vie. Il est vrai qu'on perd quelquefois sa peine en bienfaisant, et au contraire on gagne à mal faire : mais cela ne peut changer la règle de prudence, laquelle ne se rapporte qu'aux choses qui arrivent le plus souvent. Et pour moi, la maxime que j'ai le plus observée en toute la conduite de ma vie,

> a été de suivre seulement le grand chemin, et de croire que la principale finesse est de ne vouloir point user de finesse."[44]

For simplicity we may assume that Descartes has in mind something like the Prisoner's Dilemma. He then seems to argue that even if it is rational to choose the non-cooperative strategy in a single game of this type, it would be rational to cooperate in a sequence of such games. This general idea can be understood in either of two ways. In the first place an egoist might reason in the following manner: "If I appear outwardly as a helpful and altruistic person, other persons will come to like me and be concerned with my welfare, so that they will derive positive utility from my consumption. This means that they will tend to act altruistically towards me, and the net result will be that my welfare is larger than it would have been if I had acted strictly egoistically. For in general the benefits of a sacrifice to the person for whom it is made exceed the cost to the person by whom it is made, so that given equal sacrifices on all sides, the benefit of the sacrifices made for my sake will exceed the cost of the sacrifices that I make for the sake of others." This argument (which is rather similar to some ideas recently proposed by Gary Becker [45]) works quite well as long as we confine ourselves to a single individual acting in a world of other individuals whose reactions he can predict in the specified manner. If, however, the other actors are allowed to reason in the same manner, and to know that all reason in this manner, the argument collapses, because no one would have any feeling of friendship towards a person who acts altruistically only to obtain the benefits from the friendship.

In the second place we might appeal to the analysis of supergames proposed by Martin Shubik and Michael Taylor [46]. The basic idea of this approach is that if (*i*) the number of games is either a random variable or at least unknown to the actors and (*ii*) there is a discounting of the future, then it may be individually rational to choose the cooperative strategy in the Prisoner's Dilemma. Both of these conditions are very plausible in real-life situations: we do not know *when* we shall die (so the first condition is fulfilled), but we do know *that* we shall die (so the second is also fulfilled). The problem, however, is that the cooperative behaviour is not a dominant strategy in the supergame, so that for an individual to choose that strategy he would have to be very certain that the others are going to do so as well. This means that the second approach to the "rationality of altruism" runs into the opposite problem of the first one: the second approach is successful only if public and known to everyone, the first only if hidden and unsuspected. This also means that in a world with uncertainty and suspicion *neither* approach will be successful. As real-life situations have uncertainty and suspicion as rather prominent features, this conclusion is a sad one and would seem to undermine the basis for the Cartesian optimism.

Let us, however, abstract from this problem and just assume that Descartes is right in arguing (along either of the two lines distinguished above) that it is rational in the long run to help others even if clearly irrational in the short run. Returning to the comparison between Pascal and Descartes, we might ask how a person should go about it if he has decided that "le grand chemin" is the optimal course. Will the abstract calculations suggested above be sufficient to let help be forthcoming even when at the expense of the helper, or should he rather try to set up a character from which help will flow "sans violence, sans art, sans argument"? Should reason be present at each choice, or only once and for all at the choice that sets up the *hexis* from which the later choices will then follow "automatically"? Pascal would answer that the latter procedure is required, and I think he would be right. The arguments sketched above do not have such an intuitive and compelling nature that it is possible to keep them before me each time I am tempted to seek my short-term gain. The same argument holds for the traveller seeking his way out of the forest. Even if he has decided not to change direction "pour de faibles raisons", these reasons may seem more compelling when thirst and fatigue overwhelm him. The meta-rational actor, who is able to predict this possibility, would then put blinkers on so as to make himself physically unable to gather the information that lends some force to these reasons. This is a visual analogy to the strategy of putting wax in the ears, chosen by Ulysses for his men. An analogy to the strategy chosen by Ulysses for himself would be to preset a course and then in some way make oneself unable to change it. Actually neither of these strategies corresponds to the Cartesian analysis in *Les passions de l'âme*, which would rather recommend a planned change of character, so that even if one were able to gather the information and to act upon it, one would not wish to do so.

V

In this Section I shall consider the problem of inconsistent time preferences that was first raised by R.H. Strotz some twenty years ago. In the next Section I shall deal with the problem of endogenous change of preferences, that has been known to exist at least since Marshall (or Aristotle) but has received systematic discussion only the last ten years or so. These two problems are closely related, and it is possible to treat them in a unified framework, as in a recent article by Peter Hammond[47]. Nevertheless I believe that there are good reasons for following the tradition and treating the two separately. Both inconsistent time preferences and endogenous change of preferences imply that past decisions are not followed up, and they both assume that this can occur even in the absence of external influences such as persuasion, propaganda or advertising. In this sense both mechanisms are non-exogenous. Inconsistent time preferences, however, do not imply any character modification of the individual, as do the endogenous preference changes.

No doubt these changes, being endogenous, must be explained through some fixed tranformation laws that characterize the individual and that might be seen as constant higher-order character features. The change of character is merely the unfolding of a potential. Choice according to inconsistent time preferences, on the other hand, is the progressive unfolding of an *actual* and unchanging attitude towards time.

It is widely, but far from universally, agreed that for an individual the very fact of having time preferences, over and above what is justified by our knowledge that we are mortal, is irrational and perhaps immoral as well [48]. (For *societies* planning for an infinite future it would be irrational *not* to have time preferences, because otherwise consumption will be postponed indefinitely for the sake of more investment.) [49] A common view of the matter is, perhaps, the following. *Qua* rational beings (and abstracting for simplicity from the problem that we know that, but not when, we shall die) we want to allocate our welfare evenly over time. A year is a year is a year; there are no rational grounds for preferring the present over the future simply because it is present. (I am assuming here that time preferences always imply a preference for the present over the future, which disregards the very interesting problem of the miser.) On the other hand, there is the fact of weakness of will. We simply are not able to follow our rational inclinations in many cases, and the time preferences may be seen as the formal expression of this inability. Someone preferring the present over the future *must* allocate a larger amount of consumption to the present than to the future, regardless of the rational or sophisticated arguments for doing otherwise. This point will emerge as being crucially important for the discussion below of the strategy of consistent planning.

Having a preference for the present will typically lead me to *regret* my past decisions when a new present — the past future — comes along and reveals itself to be an equally worthy candidate for my attention. Now if one *regrets* a past decision, the rational thing would seem to be to *reconsider* it. If, in the past, I have decided upon an allocation of goods over (what was then) the future, and if I now come to regret the decision, would not the rational response be to reconsider the decision for what remains of the future? The problem, however, is that my only disagreement with my past decision may be with the allocation of goods between the past and the present, whereas my former and my present selves may agree upon the allocation over what is now the future. And as a matter of fact, economists have not argued that reconsidering a past decision is a way of making it more rational; rather they have argued that such reevaluation represents an additional element of irrationality, at least under the conditions that we shall now go on to make more precise.

I assume that an individual has one unit of consumption goods that shall be distributed over a number of years. In the numerical examples below I consider only three-year and four-year cases. The "years" should be interpreted as "naturally lumpy" units of time; as reflecting a non-arbitrary division of the temporal continuum. For reasons discussed below I believe it important to look at the present as something more than a point at zero distance from now. It is, phenomenologically, the period that includes the mathematical present and that is separated in some "natural" manner from other periods. Such natural divisions include the days, which are separated from each other by periods of sleep "when I do not exist"; the time from pay day to pay day ("I shall start saving at next pay day"); a calendar year ("I shall stop drinking at January 1 next year") and so on. In any given year i the time preferences up to the last year n of the period over which the goods are to be allocated, can be summed up in a utility function U = u_i (C_i, C_1... Cn). I shall restrict myself to utility functions of this general form :

$$u_1 = a\ln C_1 + b\ln C_2 + c\ln C_3 + d\ln C_4$$
$$u_2 = a\ln C_2 + b\ln C_3 + c\ln C_4$$
$$u_3 = a\ln C_3 + b\ln C_4,$$

with the obvious modification for the three-period case. These logarithmic utility functions are easy to handle mathematically, and they make economic sense in so far as they exhibit decreasing marginal utility. It is of course true that there is an infinity of other functions with the same feature, but as the logic of the argument proposed below is existential rather than universal, relying on counter examples rather than on general theorems, this restriction does not really matter.

Two more features of these utility functions should be noted. In the first place we assume that they are *constant*, in the sense that in year i the relative importance of consumption in year i compared to consumption in year i + 1 is the same as the relative importance in year j of consumption in year j compared to consumption in year j + 1, for all i and j up to and including n — 1. (This makes precise the statement above that we shall discuss the problem of time preferences independently of the problem of changing preferences.) In the second place the logarithmic utility functions permit a natural assessment of the rationality of a given allocation. For an allocation (C_1, C_2, C_3, C_4), where $\sum C_i = 1$, we define the *rational value* of the allocation as

$$u(C_1, C_2, C_3, C_4) = a\ln C_1 + a\ln C_2, + a\ln C_3 + a\ln C_4$$

which means that all years are given equal weight. The obvious candidate for a rationality index would then be the ratio of this rational value to the rational value of the rational allocation (1/4, 1/4, 1/4, 1/4), but as our only interest is in ordinal comparisons we can simplify even more and define the rationality index as

$$r(C_1, C_2, C_3, C_4) = 256 \cdot C_1C_2C_3C_4$$
$$r(C_1, C_2, C_3) = 27 \cdot C_1C_2C_3$$

for the four- and three-period cases respectively.

Given these preliminaries we are now in a position to define and discuss the notion of *consistent* time preferences. We start with some three-period examples. At the beginning of the first year the individual chooses an allocation that maximizes $u_1(C_1, C_2, C_3)$, subject to the budget constraint $\sum_i C_i = 1$. We assume that this gives the values C'_1, C'_2, C'_3. At the beginning of the second year the individual plans an allocation that maximizes $u_2(C_2, C_3)$, subject to $C_2 + C_3 = 1 - C'_1$. We assume that this gives the values C''_2 and C''_3. The consistency of the time preferences can then be defined as the requirement that $C'_2 = C''_2$; that the amount planned in the first year for the second year is the same as the amount chosen in the second year for the second year. This definition is easily extended to the n-period case, by stipulating that an individual having consistent (and constant) time preferences should never have to reconsider his past decisions, even though he may come to regret them. With utility functions having the form specified above, it is easy to show that time preferences are consistent if and only if $a/b = b/c$ in the three-period case, and $a/b = b/c = c/d$ in the four-period case. In other words, constancy and consistency of time preferences imply exponential decay of the future. For a numerical example we may take

$$u_1 = 4\ln C_1 + 2\ln C_2 + \ln C_3$$
$$u_2 = 4\ln C_2 + 2\ln C_3$$

which gives the allocation (4/7, 2/7, 1/7). This allocation is irrational, in the sense of exhibiting a preference for the present, but nevertheless consistent, in that there is no need for revision of former plans. I shall call it, therefore, the *consistently irrational allocation*. As our first example of *in*consistent time preferences we take

$$u_1 = 3\ln C_1 + 2\ln C_2 + \ln C_3$$
$$u_2 = 3\ln C_2 + 2\ln C_3$$

A person exhibiting these preferences would plan in year one the allocation (1/2, 1/3, 1/6), but when the second year actually arrives the allocation for the last two years will be reconsidered and gives (3/10, 2/10). All in all this implies that the *inconsistently irrational allocation* will be (5/10, 3/10, 2/10).

The inconsistently irrational person is not only myopic; he is myopic in an inherently contradictory manner that never permits him to stick to past decisions. Strotz assumes that such a person might nevertheless be rational enough to understand his predicament and deal with it in a strategic manner. This is not as strange as it may sound. Consider a person who has inherited a fortune and decides to spend it in the following manner. In the first year he will spend half his fortune on an enormous spree, and then he will divide the other half evenly over the rest of his life. When the second year arrives, however, he prefers — true to his (inconsistent but constant) time preferences — to spend half of the remaining half on a somewhat smaller spree in the second year, and then divide what now remains evenly over the rest of his life, and so on. (Observe that the point is *not* that he has become addicted to the "dolce vita".) It is not at all psychologically implausible that such a doubly irrational person — having time preferences, and inconsistent ones at that — might know or come to know his own character and to take precautions against his later selves so that they will not betray him. In this case the obvious way out is to buy an annuity that cannot be reconverted into cash. It is easy to see that an annuity is of no avail against consistent irrationality. Either the person buys an annuity that diminishes exponentially with time, but in that case the annuity only does what he would have done by himself (assuming consistency). Or he buys an annuity that gives him the same amount in each year, including the first, but then he is rational. Or he buys an annuity that gives him a larger amount in the first year and smaller equal amounts in each later year, but this is the inconsistently irrational case.

These remarks suggest that inconsistent irrationality may be a more frequent and more important phenomenon than consistent irrationality. E.S. Phelps and R. Pollak [50] have proposed a general mechanism whereby preferences become non-exponential and thus inconsistent. They argue that time preferences may be decomposed into two distinct discounting functions: one which gives an absolute priority to the present over all later times, which are given smaller and equal weights; and one which decays exponentially. I think this distinction is psychologically sound, because in the notion of time preferences we pack two distinct problems that can and should be separated: the absolute priority of the present and the gradual shading-off of the future. The absolute priority of the present is somewhat like my absolute priority over other persons: I am I — while they are all "out there". The shading-off is a perspectival phenomenon that admits of degrees of "out-there"-ness: the far future is like a distant relative, while the near future is more like a close one. (It is at this point of the argument that the notion of a "thick" present is required, for in continuous time it would be hard to make sense of the absolute priority of the (mathematical) present.) Of these two discounting functions the first, by itself or in conjunction with the second, is sufficient to destroy the consistency of the preferences. This provides an incentive for the individual to bind himself; while feeling incapable of curbing his present desires, he does feel capable now of setting up some machinery that will prevent him from succumbing to the future desires that have not yet achieved this absolute status. The logic of the predicament is summed up in Augustine's prayer: "Give me chastity and continence, only not yet." [51]

In the case of the inconsistent preferences defined above we can specify the *Ulysses allocation* as (1/2, 1/3, 1/6). The individual is able to bind himself to the allocation that he prefers initially, so as to achieve consistency. Now the general thesis of my essay is that such precommitment strategies are second-best solutions that could form the core of a theory of imperfect rationality, but this case might seem to constitute a counter-example. We observe, in fact, that the rationality index of the Ulysses allocation is 3/4 while that of the inconsistently irrational allocation is 81/100: the naive and myopically inconsistent person achieves a more uniform and more rational allocation over time than the sophisticated person capable of binding himself. If the sophisticated person comes to understand this, would not the supreme sophistication be to renounce at all sophistication ("la principale finesse est de ne vouloir point du tout user de finesse") and just follow one's natural bent even if this would imply the reconsideration of past decisions?

Before we answer this question it is useful to look at another sophisticated approach, the strategy of consistent planning as discussed by Strotz and refined by R. Pollak [52]. Here we assume that the sophisticated person reasons as follows. "If for the first year I choose some amount C_1 (never mind how this choice is made), then I shall have $1—C_1$ available for consumption in later years. In the absence of precommitment, and given my character as I know it, I shall allocate this amount over the last two years according to u_2. This means that I shall allocate $\frac{a}{a+b}(1-C_1)$ to the second year and $\frac{b}{a+b}(1-C_1)$ to the third. This means that only consumption sequences of the general form $\left(C_1, \frac{a}{a+b}(1-C_1), \frac{b}{a+b}(1-C_1)\right)$ can be realized. Knowing this, I should choose the best among these sequences, *i.e.* I should choose C_1 so as to maximize $a\ln C_1 + b\ln\left(\frac{a}{a+b}(1-C_1)\right) + c\ln\left(\frac{b}{a+b}(1-C_1)\right)$.

For the case discussed above (a = 3, b = 2, c = 1) the ensuing *allocation of consistent planning* is (18/53, 21/53, 14/53), with a rationality index about 0.96.

This is a rather surprising result. The allocation of consistent planning is quite close to the allocation of rational planning, involving slightly more in the first year, considerably more in the second and substantially less in the third. Note, however, that even the third year fares better on the strategy of consistent planning compared to its fate under the inconsistently irrational allocation and the Ulysses allocation. There is, I think, a paradox in this rational character of the strategy of consistent planning, for if someone were able (*i.e.* had the willpower) to follow this strategy, then he should surely also be able to follow the rational strategy. If the solution is feasible, the problem to which it is a solution does not exist. To this objection a possible counter-argument might be as follows. The strategy of consistent planning is a response to the problem of *inconsistently* irrational preferences and not to the problem of inconsistently *irrational* preferences. The issue is how to cope with my predictable lack of ability to act upon my present intentions, and not how to cope with the irrational component of these intentions themselves. This line of argument might also be used to answer the query raised above concerning the Ulysses strategy, whether naiveté could be more sophisticated than sophistication. The Ulysses strategy is a precaution against inconsistency, not against irrationality; in fact it achieves consistency at the cost at an even larger departure from rationality, at least in the numerical example given above.

Before proposing an answer to the counter-argument it will be useful to have a larger variety of examples. I list here below some utility functions and the allocations (with the rationality indices) that follow for each of them, including, for ease of comparison, the case discussed above.

		Rational allocation	Ulysses allocation	Consistent planning allocation	Inconsistent irrational allocation
A	$u_1=3\ln C_1+2\ln C_2+\ln C_3$ $u_2=3\ln C_2+2\ln C_3$	1/3,1/3,1/3; r=1	1/2,1/3,1/6; r=0.75	18/53,21/53, 14/53;r=0.96	1/2,3/10,1/5 r=0.81
B	$u_1=3\ln C_1+\ln C_2+\ln C_3$ $u_2=3\ln C_2+\ln C_3$	1/3,1/3,1/3; r=1	3/5,1/5,1/5; r=0.65	9/25;12/25, 4/25;r=0.75	3/5,6/25,4/25 r=0.62
C	$u_1=4\ln C_1+3\ln C_2+2\ln C_3$ $u_2=4\ln C_2+3\ln C_3$	1/3,1/3,1/3; r=1	4/9,1/3,2/9; r=0.89	48/167,68/167, 51/167;r=0.97	4/9,20/63, 5/21;r=0.91
D	$u_1=4\ln C_1+2\ln C_2+2\ln C_3$ $u_2=4\ln C_2+2\ln C_3$	1/3,1/3,1/3; r=1	1/2,1/4,1/4; r=0.84	4/13,6/13, 3/13;r=0.88	1/2,1/3,1/6 r=0.75
E	$u_1=3\ln C_1+2\ln C_2+\ln C_3$ $u_2=3\ln C_2+2\ln C_3+\ln C_4$ $u_3=3\ln C_3+2\ln C_4$	1/4,1/4,1/4, 1/4;r=1	1/2,1/3,1/6,0; r=0	96/361,90/361, 105/361,70/361; r=0.96	1/2,1/4,3/20 1/10;r=0.48
F	$u_1=2\ln C_1+\ln C_2+\ln C_3+\ln C_4$ $u_2=2\ln C_2+\ln C_3+\ln C_4$ $u_3=2\ln C_3+\ln C_4$	1/4,1/4,1/4, 1/4;r=1	2/5,1/5,1/5, 1/5;r=0.819	8/47,12/47, 18/47,9/47; r= 0.816	2/5,3/10,1/5 1/10;r=0.61
G	$u_1=4\ln C_1+2\ln C_2+\ln C_3$ $u_2=4\ln C_2+2\ln C_3$	1/3,1/3,1/3; r=1	4/7,2/7,1/7; r=0.63	4/7,2/7,1/7; r=0.63	4/7,2/7,1/7; r=0.63

Before commenting upon the table, cases E and F require a few words of explanation. When we go from the three-period to the four-period case, we have a choice between two interpretations of the strategy of consistent planning. Should we interpret this strategy as a sophisticated decision in year one trying to make the best out of the naive decisions that are predicted for years two, three and four — or should we impute to the decision-maker of year one the assumption that the year-two decision will also be a (lower-level) strategy of consistent planning? Pollak opts for the latter possibility, and, as he observes himself in another context (see note 50), it is indeed reasonable to assume (*i*) that sophistication once acquired will never be forgot and (*ii*) that a sophisticated decision-maker is aware of (*i*).

Case G is the example of consistent preferences used above; case A the numerical example of inconsistent preferences used for the preliminary discussion. Cases B, D and F all correspond to the "go-on-a-spree-in-the-first-year-and-then-divide-evenly" case discussed informally above. Cases C and E are included for some additional variety. We observe that in all cases but case F the strategy of consistent planning is the most rational one. The Ulysses allocation is the most rational one in case F. The Ulysses allocation is sometimes less rational than the inconsistently irrational allocation, sometimes more rational. When it is the more rational, consistency and rationality go hand in hand; otherwise the two are opposed to each other, at least if we exclude the strategy of consistent planning, *as I think we should do.* I believe that the strongest argument against the usefulness and indeed the meaningfulness of the strategy of consistent planning is provided by an examination of case F. Here we have a decision-maker with quite strong preferences for the present, which according to our interpretation of the time preferences means that he simply does not have the willpower to achieve rationality. Nevertheless we are asked to believe that such a person could adopt the strategy of consistent planning, which in this case implies *less* consumption in the first year — the year of the decision — than in all subsequent years. It is as if the inconsistent spendthrift could solve his inconsistency problem by turning into a miser; as if the inconsistent preferences for the present over the future can be made consistent by turning them into preferences for the future over the present. The same feature is found in case C, and here the consistent planning allocation is also the most rational one. The point, I believe, is the following: a person deficient in willpower as well as in consistency might very well decide upon the strategy of consistent planning in order to minimize the harm done to him by his future selves, but his lack of willpower would prevent him from carrying out that decision. The strategy of consistent planning could be a solution to the problem of inconsistency only if it was no more rational (*i.e.* required no more willpower) than the inconsistently irrational allocation, but this is not the case in any of the examples in the table.

The only answer I can imagine to this response to the counter-argument is that the year one allocation that is out of reach of my willpower when described as part of a rational plan, might come to look more acceptable when *redescribed* as part of a package solution that takes some account of my present preferences u_1. This suggestion would have a family resemblance to Ainslie's notion of private side bets, which also turn upon a redefinition or a rearrangement of my inner space. The suggestion, that is, would be that for some individuals it may be impossible to act rationally if the motive for the rational behaviour *is* to act rationally, whereas they might achieve rationality if obtained as a by-product of actions undertaken for the sake of consistency. This would not be a totally implausible notion. As mentioned in Section II above, there are many cases when you simultaneously want to achieve x and yet do not want to undertake any actions *in order to* achieve x.

I believe to have shown that, with the possible exception discussed in the preceding paragraph, the strategy of consistent planning is not a feasible option in the face of inconsistency and irrationality. In some cases the Ulysses allocation can achieve both consistency and the best approximation to rationality, but in other cases it achieves consistency at the cost of rationality. The choice between the Ulysses allocation and the inconsistently irrational allocation then depends upon which desire is the strongest, the desire for consistency (the wish to impose my decision upon my later selves) or the desire for rationality (my concern with the welfare of my later selves). I believe that in such cases of conflict moral reasons should prevail, so that the hyper-sophisticated decision to abandon sophistication would be the correct one. To bind oneself makes for consistency, but the price paid may be too high. Situations such as cases A, C and E in the table can then be said to constitute genuine exceptions to the thesis that imperfect rationality is best achieved through precommitment; unless, of course, one is willing to work on the preferences themselves. I now turn to this case.

VI

A (very) general theory of human action can be sketched as follows. When we ask why a person in a given situation behaves in one way rather than in another, the answer is to see his action as the result of two successive filtering processes. The first has the effect of limiting the set of abstractly possible actions to the *feasible set*, *i.e.* the set of actions that satisfy simultaneously a number of physical, technical, economic and (perhaps) politico-legal constraints. The second has the effect of singling out one member of the feasible set as the action which is to be carried out. From this theory we can immediately conclude that the act of binding oneself, as explained in Section II above, can either be directed towards a change in the feasible set or towards a change in the mechanism that picks out the feasible alternative to be realized. In the preceding Section we assumed that the Ulysses strategy implied an induced change in the feasible set; here we shall explore the other possibility. We shall also assume that at any given moment of time the singling out of one member of the feasible set is effectuated through rational choice, *i.e.* that the individual in question chooses the feasible alternative that is the most preferred according to his (consistent) preferences [53]. This implies that precommitment must take the form of acting upon the preferences, as recommended by Descartes in *Les passions de l'âme*.

This problem is virtually unexplored in the literature. There is a small (but very rapidly growing) body of literature on the notion of endogenously changing preferences [54], but very rarely is it assumed that an individual can take an active or strategic attitude towards his own preferences. Rather he is seen as the passive vehicle of preferences that change according to some logic that he does not himself understand. Most authors mention in passing that the individual can take his precautions against such change if he comes to understand the mechanism by which it is brought about; resistance to habit-forming drugs is the example usually cited. The implications of this approach, however, are rarely worked out, and the main emphasis usually is upon the analytical and the manipulative aspects: how do preferences change, and how can this be exploited by other people? I shall also discuss these latter aspects, and then try to relate them to the main theme of this essay.

I shall start with a discussion of C.C. von Weiszäcker's seminal paper on "Endogenous change of tastes" [55]. For simplicity he assumes (*i*) that current preferences depend only upon consumption in the immediately preceding period and (*ii*) that the income of the consumer is to be allocated over two goods only. It has been shown recently that the second assumption is crucial for the results obtained by Weiszäcker, and that they do not generalize to the n-commodity case [56]. I shall disregard this problem, however, as I think there are philosophically important questions to be discussed even in the special case. The first assumption may seem rather strange, as it excludes the notion of cumulative character formation which most people have in mind when they discuss endogenous preference changes. It is probable, however, that the introduction of the whole consumption history as a variable would complicate the analysis so as to make it virtually unmanageable, and that at the present stage of research the strategy chosen by von Weiszäcker is the best one.

Economists, when studying the consumer's choice, use more or less interchangeably three distinct concepts: preferences, utility functions and demand functions. Given various conditions, which we shall assume fulfilled, any two of these concepts may be deduced or constructed from the third. Let us suppose that the consumer at time t has a preference structure P_t, represented by a utility function u_t, from which we can derive a demand function f_t which to each set of prices and income correlates a certain (two-component) commodity vector chosen by the consumer: $q_t = f_t\ (p, m)$. We then assume that present demand is a function of demand and consumption in the immediately preceding period: $f_t = F\ (f_{t-1}, p_{t-1})$, where F is some functional that represents the constant character traits underlying the changing preferences. More specifically we assume that F is such that there exists a g such that $q_t = g(p, m, q_{t-1})$. This latter form is the only one used by Weiszäcker. Successive applications of the function g generate a sequence q_t of commodity bundles, which under certain conditions (whose economic interpretation is not clear) is shown to converge to a point $q = h(p, m)$ that is independent of the initial bundle. The relation h may be formally interpreted as a demand function, as to each set of prices and income it correlates a certain commodity bundle. To this demand function there also corresponds (given certain controversial assumptions) an "as-if" utility function U, in the sense that we can predict the bundle $q = h(p, m)$ if we assume that the consumer is maximizing U, given p and m. Alternatively we may construct the "long-term preference structure" with the property that xPy if and only if $U(x) > U(y)$.

Let us return to the "instantaneous" preferences P_t. There really is no need to index these by *time*, because the only relevant difference between P_{t_1} and P_{t_2} is the consumption in the immediately preceding periods t_1-1 and t_2-1. If consumption in these periods was the same, then the instantaneous preferences in the following periods will also be the same. We may therefore more usefully index preferences by consumption in the preceding period, writing "xP(y)z" for "x is preferred to z, given that consumption in the immediately preceding period was y". Weiszäcker now proves the following theorem, valid under conditions that verbally may be expressed as inertia or conservatism of the consumer:

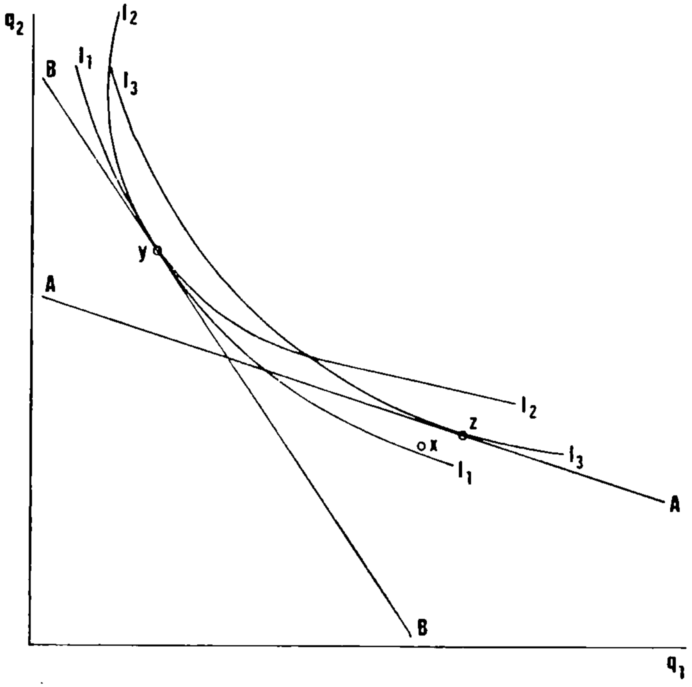
> Given two commodity vectors x and y, then xPy if and only if there exists a sequence of vectors $r_1 \ldots r_n$ (where $r_n = x$) such that $r_1P(y)y$, $r_2P(r_1)\,r_1$, $r_3P(r_2)\,r_2 \ldots r_nP(r_{n-1})r_{n-1}$.

The theorem says, in other words, that x is better than y according to my long-run preferences if and only if there exists a path from y to x such that each step in the path is an improvement according to the instantaneous preferences at the beginning of that step. *If* the long-term preferences are interpreted as my real or as my rational preferences, then the theorem implies that even if there is a conflict between my short-term preferences and my real ones, in the sense that yP(y)x, this conflict can always be overcome by the use of indirect strategies. It is a childhood fantasy come true: a walk in a landscape where from any given point you can always come to any higher point without ever having to go uphill, because the landscape contours change as a function of your path.

Weiszäcker then gives an ingenious and controversial example of this approach at work. In the figure below the amounts of goods q_1 and q_2 are measured along the axes. Line BB represents the budget line of a farmer, AA the budget line of an industrial worker. Initially the person in question is a farmer enjoying the consumption bundle y. I_1I_1 is the long-run indifference curve passing through y and I_2I_2 the short-run indifference curve corresponding to the preferences at y. z is the long-term equilibrium consumption point if the farmer were to move to the city. The point x represents a consumption bundle having the following properties :

(*i*) zPx
(*ii*) xPy
(*iii*) yP(y)x

If the farmer consumes x when living in the city, there will be a surplus available for taxation. Now city life (both in version z and a fortiori in version



x) does not look very attractive from the point of view of the farmer, even though (in both versions) it is superior according to his long-run preferences. The theorem tells us, however, that there exists a sequence of points r_i that can bring him painlessly from y to x. (The reason why we look at the transition to x rather than to z will become clear in a moment.) The snag, however, is that as the curves are drawn, the initial part of the sequence r_i must lie above both budget lines, because all r_i must lie above I_1I_1. The latter statement is justified as follows. By definition each r_i satisfies the condition of the theorem that there exists a sequence of short-term improvements from x to r_i (viz. r_1, r_2... r_{i-1}), and according to the theorem this implies that r_iPy for all r_i, which again means that all r_i must lie on long-term indifference curves higher than I_1I_1. Weiszäcker proposes the following solution to this problem: the government could subsidize the intermediate points r_i that lie above AA; when an r_i falls below AA the government transfers the farmer to the city, gives him income permitting him to buy x and retains the difference between z and x for itself in order to get back the money spent on subsidies. The farmer is better off (because xPy) and the government makes a profit because the subsidies are given only for a limited period of time whereas the income from taxation can be perpetuated indefinitely (through the descendants of the farmer). Thus all is for the best in this Pareto-optimal world.

Or is it? We have just discussed the case where an external agency could *manipulate* the individual with endogenously changing preferences, by putting him in a sequence of situations that will bring him smoothly and painlessly into a certain state that he could not have achieved by himself. If this is to be ethically acceptable, we must assume that the individual — if rational — would have done the same himself given the same knowledge about the causal process underlying the preference change. We must assume, that is, that he would have applied for a loan to bring the intermediate points within his feasible set, planning to repay the loan when the final state is realized. I would not contest that this may happen in some cases; I believe in particular that the choice of a university education may be seen in this perspective. On the other hand I am unable to see that it would be irrational for an individual to refuse the offer of a subsidy, or to refrain from applying for a loan. If it could be rational to refuse such an offer, then I think it would be unethical to manipulate the individual through such offers.

This point merits some elaboration. We may distinguish between coercion, seduction, persuasion and voluntary engagement by looking at the relation between preferences before the fact and preferences after the fact. Coercion and voluntary engagement are the extreme ends of the spectrum. Coercion takes place where an individual initially prefers x over y, and continues to do so even when someone coerces him into doing y. Voluntary engagement means that the individual initially prefers y over x, and does y for that reason. (He may come to prefer x to y after the fact, but this is irrelevant in the present context.) Seduction occurs when an individual initially prefers x over y, but comes to prefer y over x once he has been coerced into doing y. Persuasion is the case discussed by Weiszäcker, where an individual is led by a series of short-term improvements into preferring y over x, even if initially he preferred x over y. No one will contest that coercion and seduction are bad, and voluntary engagement good. I would argue, however, that many cases of persuasion are so similar to seduction as to be unacceptable for the same reasons. Imagine, for example, the following situation. I am a non-smoker, and I do not like the cheap tobacco that is within my budget possibilities. If, however, someone were to subsidize my buying better tobacco, my craving for nicotine would become so great that after some time (when the subsidies are withdrawn) I would gladly accept the cheap tobacco which I formerly rejected. Are such "introduction offers" always ethically justifiable, and is it irrational to refuse them? I believe most people would agree that both questions are to be answered in the negative. The same arguments would justify the farmer refusing the subsidies, saying that he prefers his present preferences even if they do not coincide with the long-run optimum.

The last remark reminds us of the fact that the persuasion discussed by Weiszäcker is not just any kind of persuasion; it is persuasion "for my own good" in the sense of bringing me onto my "real" or "long-run" optimum. Weiszäcker might argue, therefore, that even if many cases of persuasion can be assimilated to seduction, this analogy breaks down in the kind of cases he is considering. The problem with his (putative) answer is that the long-run stability of tastes cannot provide an argument for their morality or rationality. I can see no contradiction in the notion of an addiction that at first grows rapidly and then converges to some large but finite amount. This, presumably, might express my long-term preference in the technical sense that we are discussing, but I can see no reason whatsoever why the individual should addict himself or accept offers that would lead to addiction [67].

The decision not to addict oneself is an act of omission rather than of commission, and so is excluded (by criterion (*v*) in Section II above) from the set of acts of precommitment. On the other hand the decisions to addict oneself (*e.g.* to good music) or to prevent oneself from offers that lead to addiction would count as examples of binding oneself. The discussion of the "paternalistic state" may be seen in the light of these notions. To the extent that bans on cigarette advertisements stem from the actual or potential consumer themselves, who want to protect themselves from the Sirens of publicity, one should not talk about paternalism. The same holds for the introduction of one television-free day a week in order to stimulate political or cultural activities, or the obligatory use of safety-belts. It goes without saying, I think, that the individuals are free to bind themselves (through laws), to bind themselves (with safety belts), or to protect their "deeper values" against the more impulsive ones. The hard problems arise when these choices are imposed upon the individuals, against their *ex ante* (and perhaps) against their *ex post* preferences. I think it fairly clear that the government has not the right to screen the public from television one day of the week if the public does not want to be screened. It is much less obvious that the government should not have the right to make the use of safety belts obligatory, for here the state is asked to pay the bill when the accident arrives. The question, in a nutshell, is whether the welfare state implies or justifies the paternalistic state. Some people would say yes, and their aversion for paternalism would make them reject the welfare state. Other people might say yes and accept paternalism because they accept the welfare state. Still other people might say no, and accept the welfare state without paternalism. To the extent that this reflects the preferences of the individuals that make up society (and I believe it does), this third solution should be preferred. If people do not want to bind themselves. then they should not be made to choose between being bound or getting no help if they hurt themselves as a result of not being bound. Here, of course, it is crucial that there should be some correlation between the extent to which individuals engage in risky activities and their contribution to social security funds; otherwise the refusal to bind oneself would not be an expression of spontaneity but an attempt to operate as a free rider.

VII

Among the many striking observations offered by Ainslie is the evidence that animals are able to use strategies of precommitment. Two examples are the following:

> "It is well known that pigeons will peck a key that has been associated with food, even when not pecking leads to much greater reward. Ainslie found that pigeons would regularly peck a red key for 2-sec access to food where not pecking led to 4-sec access to food beginning 3 sec later. If the key lit up green 12 sec before it was due to light up red, some of the subjects came to peck it on a majority of trials when its only effect was to prevent the key from lighting up red later in the trial [...] Subjects apparently learned to peck the key when it was green only if this forestalled the temporary attractiveness of the red key." [58]

> "The tendency of hungry pigeons to actively avoid certain opportunities to get food seems inconsistent with the orthodox concept of reward. If the fixed number of pecks on a key required for a single food reward is high but not so high that the pecking ceases (between 50 and 2000), pigeons will peck a second key whose only effect is to make the original key unavailable for a period of time (a time-out) [..] the subjects that sought time-outs from the opportunity to get food were those that presumably faced close choices between resting and working for poor reward. In such a situation their choice could be expected to vacillate between the alternatives. If we suppose that ambivalent behaviour is less rewarding than that of sticking to either alternative, the subjects might seek a device that bound their future behaviour to one or the other, or at least reduced the frequency of change." [59]

These unexpected findings might seem to conflict with the main thesis of the present essay, which is that men are neither angels (*i.e.* fully rational) nor animals (*i.e.* essentially myopic), but that they are imperfectly rational creatures able to deal strategically with their own myopia. Do the results reported by Ainslie imply that men should be debased to the level of animals, or animals elevated to the status of men? In order to present my argument for the (negative) answer to this question, I shall have to draw upon some distinctions developed in a different context [60]. I argue there (*i*) that the capacity for global maximization is a uniquely distinguishing feature of man, and that (*ii*) the examples of global maximization (*e.g.* the use of indirect strategies such as path interception in predator behaviour) found in the animal realm evolve by accident only. Man has a *general* capacity for global maximization that he can deploy in qualitatively novel and unprecedented situations, whereas the scattered examples of such behaviour in animals are all highly situation-specific.

The argument of the present essay makes a case for a more complex form of global maximization, involving the use of indirect strategies (*i.e.* precommitment) in order to enable us to use indirect strategies (such as investment). The capacity for imperfectly rational behaviour in this sense is also a *general* capacity; the strategy of binding oneself can be applied to a variety of situations and can exploit an extremely wide range of mechanisms. The fact that pigeons can be *conditioned* to bind themselves in situations that are highly *artificial* and very *specific*, should not be interpreted in the same way as the *spontaneous* use of precommitment in a very *broad* set of *natural* settings. And even if one could demonstrate that animals use precommitment behaviour in real-life situations, the explanation of this would have to be sought in evolutionary mechanisms such as preadaptation, genetic drift or pleiotropy; the appeal to a generalized capacity for binding oneself must be reserved for man.

The point is that in human behaviour there is a *presumption of rationality* such that global maximization or strategies of precommitment are natural forms of behaviour that do not require any further explanation; rather an explanation is called for when deviations from (perfect or imperfect) rationality is observed. In the animal realm the general mechanism of natural selection creates a *presumption of myopia*, and any given case of short-term sacrifice or precommitment behaviour would require a separate explanation. To the extent that animals could be shown to behave spontaneously in one of these future-directed manners, I think we would ascribe to them mental experiences in the same sense in which human beings have mental experiences. The generalized capacity for (non-stereotyped) global maximization or strategies of precommitment presupposes an inner space where the possible states are represented. For future consequences to make a difference for present choices, these consequences must somehow be present to consciousness *and be present as unrealized and merely possible;* this, I submit, is evidence of a mental life in a very strong sense. I would not exclude that mental experiences in this sense could be found in the higher vertebrates, such as "Imo the monkey genius" who was capable of using an indirect strategy in a qualitatively novel context [61], but for Ainslie's pigeons we lack the element of spontaneity that is required.

VIII

The quandary of Ulysses also enters into any discussion of politics. Here I shall deal with two cases that seem to be especially important: the problem of democracy and the problem of the capitalist state.

A direct democracy — either in the sense that all citizens vote on all issues instead of electing representatives, or in the sense that representatives can be recalled at any time — will tend toward zig-zag policies and towards constant reevaluation of past plans; it will be incontinent, vacillating and inefficient. Hegel is only one of the many political philosophers who have argued that this total freedom passes over into total unfreedom, both in the conceptual sense that liberty unlimited *is* slavery and in the causal sense that the anarchy of total liberty tends to prepare the grounds for a dictator. I shall briefly indicate how classical and modern democracies have evolved different ways of coping with this problem.

In classical Athens all important decisions were taken in the assembly of citizens, meeting at least 40 times a year. The assembly could reconsider its past decisions at any time, which would have left it an easy prey to demagogues or other manipulations had not certain institutional safeguards been established. M.I. Finley has singled out two institutions in this context. The first is ostracism, which in practice was the right to banish well-known demagogues. Unlike Ulysses the assembly could not always prevent itself from acting upon the song of the Sirens (but see the second institution below); it could, however, bar itself from hearing the song. The second institution was called *graphe paranomon*, "whereby a man could be indicted and tried for making an 'illegal proposal in the Assembly' [...] *even if that proposal had been passed by the Assembly* "[62]. Finley observes that this arrangement differs from the corresponding safeguards in modern democracies: "There lay the logic in the *graphe paranomon*, in the notion that by this procedure the demos was taking a second look at the proposal, rather than that one branch of government, the judiciary, was reviewing the actions of another branch, the legislature "[63]. He also suggests a contrast between the Greek approach to stabilization of politics and the political scientists who have argued that some degree of *apathy* is a necessary prerequisite for the viability of democracy [64]. I shall now argue that both of his observations, even if partly correct, are misleading in important respects.

In the first place Finley seems to have underestimated the subtlety of the *graphe paranomon*. On his analysis it could be compared to the laws that have recently been passed in several countries and that permit consumers to annul instalment contracts within a period of ten days. (Strotz argues [65] that buying on instalment is a device for precommitment, but as in the manic-depressive case discussed in Section II it may be reasonable to precommit yourself against precommitments). It seems, however, that a better comparison would be if the salesman not only ran the risk of having his contract annulled, but actually incurred a fine should it be annulled. Anyone wanting to exploit the impulsiveness of the people would then know that he runs the risk of being punished at the spur of a later impulse or when the impulse gives way to reason, and this must act as a deterrent against such exploitation. Thus the *graphe paranomon* was a stabilizing device not only in the sense of permitting reason to regain the upper hand, but in the sense of reducing the probability that it should lose out to passion in the first place.

Secondly the Greek institutions and the phenomenon of apathy can hardly be set on a par as stabilizing mechanisms. It is absurd, I think, to suggest that a people could *deliberately* opt for apathy in order to protect itself against its tendency towards excessive meddling. Apathy may or may not have this *effect;* it may or (much more probably) may not have this *function* [66]; but it certainly could never be set up with this *intention.* On the other hand, the phenomena of ostracism and *graphe paranomon* could very well have been set up in order to keep democracy within the limits required for efficiency. And according to criterion (*i*) in Section II above this is required if these institutions are to be subsumed under the heading of precommitment. The mere argument that these institutions have this stabilizing effect or (a more obscure suggestion) function would not be sufficient for this characterization. Rationality is indissolubly linked to intentionality, and this holds for imperfect rationality as well as for perfect rationality. Maximization without intentionality (as in natural selection) can only give us a tendency to seek local optima, whereas I am arguing that the capacity for non-myopic behaviour is an essential feature of human rationality.

In modern democracies a number of institutions can be interpreted as devices for precommitment. I have already observed in Section IV above that central banks can be seen as the repository of reason against the short-term claims of passion, an argument that has been developed in detail by F. Sejersted [67]. For classical liberalist policy the exchange rate and the price level were seen as crucial parameters that under no circumstances should be transformed into political control variables. Mercantilism (and present-day neomercantilism) have rather stressed the need for current adjustments to the changing environment: should Ulysses let himself be tied to the mast if he knew that the shallow waters around the Sirens' island were too difficult for anyone but him to master [68]? Other institutions that have been accorded a similar autonomy for similar reasons include the foreign ministries of many countries and the BBC model of broadcasting (as distinct from the ORTF model). For all these institutions it is possible to identify, with varying degrees of precision, the act of abdication whereby politicians have decided that certain values are too important, or certain tools too dangerous, to be subject to the current control of the politicians. To remove monetary policy, foreign policy or broadcasting from the political sphere is itself a political act.

At another level the system of periodic elections can be interpreted in the same perspective, especially when the government does not have the right to dissolve parliament and order new elections. (The democratic distrust of plebicites is an expression of the same attitude.) In this interpretation periodic elections are the *electorate*'s method for binding itself and for protecting itself against its own impulsiveness. We observed above, however, that for the *politicians* the system of periodic elections makes for a permanent temptation to lump the unpopular measures in the beginning of the electoral period and the popular ones towards the end, with suboptimal consequences. We also noted Lindbeck's suggestion that the politicians could bind themselves so as to avoid this temptation by randomly spaced elections, which would be an answer to the question "Who shall guard the guardians?". (Randomly spaced elections could also, of course, be the electorate's device for simultaneously binding itself and the politicians.) On the other hand this device would certainly be inefficient in other and probably more important respects, to the discussion of which I now proceed.

"The rule of law" has been a crucial element in modern Western democracies, especially during the 19th century. Max Weber and many others [69] have stressed the importance of a stable institutional context for a growing economy. If long-term planning, saving and investment are to be possible, the economic agents must be able to foresee future decisions by the bureaucracy. This requirement actually has two components. The law must be *unambiguous*, reducing to a minimum the need for discretionary interpretation. (This element was lacking, for example, in ancient Chinese law.) [70] The law must also be *constant* over time, even if short-term modifications might seem desirable. (This element is lacking in neomercantilist economic policy as currently practiced in many countries.) These formal requirements are in many respects even more important than the need for just laws, because if you can predict the decisions of the court you can take precautionary measures that will protect you from unjust laws [71]. In many cases the stability making for long-term optimality is also more important than short-term efficiency [72].

F. Sejersted [73] has argued that the rule of law is a solution to the problems that arise within two other political systems: the despotism that preceded parliamentary democracy and the direct democracy that is sometimes held up as an alternative to it. The latter systems are both based upon the right to current intervention in all matters, whereas representative democracy is founded upon a set of stable institutions that cannot be undone at will once established. A crucial notion in this connection is the function of the *constituent assembly*, that lays down the ground rules to be followed by all later generations. Only the constituent assembly really is a political actor, in the sense of "la politique politisante"; all later generations are restricted to "la politique politisée" or the day-to-day enactment of the ground rules. The nation can bind "itself" (a controversial notion) through the constituent assembly, by entrusting certain powers of decision to the judiciary branch, by requiring that the ground rules can only be changed by 2/3 or 4/5 majority, and so on.

The paradox of democracy can be formulated in this way: each generation wants to be free to bind its successors, while not being bound by its predecessors. This contradiction [74] has a structure similar to what has been called the central contradiction of capitalism: each capitalist wants low wages for his own workers (this makes for high profits) and high wages for all other workers (this makes for high demand). In both cases it is possible for any given generation (any given capitalist) to have its cake and eat it, but all generations (all capitalists) cannot simultaneously achieve this goal. The link to the problem of inconsistent time preferences (Section V above) should be obvious. The equivalent of the inconsistently irrational strategy could perhaps be found in recent Chinese politics, especially in the revolutionary phases where "the plan loses significance in that it is constantly revised" [75]. The Ulysses strategy is to precommit later generations by laying down a constitution including clauses that prevent its being changed. The constituent assembly has a unique and privileged character, not by right but by historical accident. In exceptional and unpredictable historical situations, representativity of persons and legitimacy of voting methods are decided on the spot; the drastic break with the past makes the assembly free to bind the future.

It is interesting that several political thinkers [76] have discussed the possibility of a periodic constituent assembly, so that everyone once in his life should be allowed to have a say in the most basic problems of society rather than being under the tutelage of past generations. This intention, while laudable, seems to harbour an irreducible contradiction. Let us assume, for example, that the constituent assembly has laid down (*i*) that there shall be constituent assemblies at 30-year intervals and (*ii*) that changes in the constitution between these assemblies must have a majority of two thirds. We then have to ask which majority is needed in the later constituent assemblies and who should decide it. If the original constituent assembly were to lay down the voting method for later assemblies, this would mean that it *was* privileged after all. If the later assemblies themselves are to decide this question, we immediately have an infinite regress, for by which methods shall they choose the voting method itself? The reason why this infinite regress does not (necessarily) arise in the original constituent assembly is the exceptional and charismatic character of the group, having its legitimation in objective circumstances outside itself. These circumstances make for a unity and a unanimity in the face of which procedural questions evaporate. (This is, of course, an idealized story, but not, I think, an imaginary one.) The implication of this analysis is that later generations have no obligation to feel bound by their predecessors, but neither do they have any legitimate right to bind their successors. *The constitution remains legitimate only because all alternatives lack legitimacy.* If in the constituent assembly 70 % voted for a required majority of two thirds to change the constitution whereas today 60 % is for a required simple majority only, this is not a legitimate basis for change.

We can take a closer look at this problem. At any given moment of time we can assume that there is a well-defined percentage $y = f(x)$ of the assembly that wants the majority required for a change in the constitution to be ***at least*** x %. Presumably y is a continuous and decreasing function of x. If $f(50) > 50$ and $f(100) < 100$, there must then be some fixed point x between 50 and 100 % such that x % of the assembly wants the required majority to be at least x %. One could then argue that this x should be incorporated into the constitution itself, as being the largest self-supporting percentage. One could argue, moreover, that x be currently modified in order to take account of changes in people's attitude towards the relative importance of democracy and stability. I believe, however, that such arguments are self-defeating. It would be pointless to bind future generations if they could unbind themselves in this manner. The only way out, perhaps, would be to incorporate a clause in the constitution that could not be modified by any generation, viz. that if at time t the percentage required for changes in the constitution is itself changed according to the above procedure, then this change should not come into effect before t + 30. Only in this way could one be certain that the percentage f(x) really reflects people's attitude towards democracy rather than being a rationalization of some goal (*i.e.* some constitutional change) they want to achieve *now*. Assuming that the tendency over time is towards lower values of x, this could give quite acceptable results. If, however, the state of the opinion vacillates between fundamentalism and radicalism in the interpretation of constitutional law, this procedure could have the strange feature of placing a conservative generation in a situation where it has too much leeway for constitutional change; it would regret not being more tightly bound to the mast. Further analysis of the situation would reveal further paradoxes: a liberal generation would want to enact illiberal measures in order to prevent the illiberal generation that is expected to follow it from enacting illiberal measures to bind the liberal generation that is expected to follow at one further remove, and so on. Once again we can only conclude that there is an inherent instability in all alternatives to the constitution handed down from the past, which then retains its legitimacy *faute de mieux*.

The second problem to be discussed in this Section is the nature of the capitalist state. It is a well-known fact that during the heyday of classical capitalism the political power was not in the hand of the capitalist class. Should this be interpreted as the capitalists' *abdication from power*, and if so in which sense and for which purpose? The analysis will mainly be conducted with reference to the Marxist theory of the capitalist state, but some additional authors will also be cited.

On the most general level (as found above all in the *German Ideology*) Marx's theory of the state can be exposed in the following manner. Classical political theory from Hobbes onwards had assumed that the state was a tool for realizing the cooperative solution to a Prisoner's Dilemma *played by all against all* [77]. Marx rejected this approach as being excessively general. Rather than looking at society as being composed of identical atoms having structurally identical (even if substantively opposed) interests, he argued that society should be decomposed into (at least) two distinct classes, each of which has its internal Prisoner's Dilemma. He argued, moreover, that the capitalist state has the double task of permitting the cooperative solution to the Prisoner's Dilemma played by the capitalists against each other and of preventing the cooperative solution (organization and solidarity) to emerge in the Prisoner's Dilemma played by the workers against each other [78]. (In addition there might be some tasks that are in the interest of both classes, such as the provision of public goods, but Marx argued [79] that the development of capitalism would lead to these being increasingly assumed by private enterprise.) The state's apparatus for repressing the workers is not relevant for our purposes, so we shall limit out attention to the capitalist state as a tool for protecting the interest of the capitalist class as a whole against the interests of its individual members and for protecting the long-term interest of the class against its short-term interests.

An example where the class interest conflicts with the individual interests is provided by the English Factory Acts, which, far from being a check on capitalist greed, were:

> "...the negative expression of the same greed. These acts curb the passion of capital for a limitless draining of labourpower, by forcibly limiting the working-day by state regulations, made by a state that is ruled by capitalist and landlord. Apart from the working-class movement that daily grew more threatening, the limiting of factory labour was dictated by the same necessity which spread guana over the English fields. The same blind eagerness for plunder that in one case exhausted the soil, had in the other torn up by the roots the living force of the nation." [80]

Such cases are relatively simple to understand. The use of legal sanctions to curb the — counterproductive — greed for profit is an instance of the general metod of public side bets which is probably the most important single device for precommitment. Other purported cases of capitalist abdication are much more complex and ambiguous. In the case just cited we are talking about a state ruled by *capitalists* (though Marx adds "and landlord") enforcing measures for the sake of the capitalist class as a whole. In other and much more important texts Marx recognized that the state is *not* a capitalist state in the direct sense of being ruled by capitalists; on the contrary the political power was firmly in the hands of the aristocracy (in England) or of a caesarist regime (in France). This separation of the economic and political power quite obviously creates a problem for a theory where one basic statement is that "The executive of the modern state is but a committee for managing the common affairs of the bourgeoisie" [81]. It is to Marx's credit that in his concrete historical analyses he dropped the "nothing but"-ism of his general theory; nevertheless it remains to be shown that the modern state — besides being many other things — *also* is a committee for managing the common affairs of the bourgeoisie.

Marx adopted two distinct lines of arguments in order to prove this for the English and the French cases respectively. In his articles for the *New York Daily Tribune* on English politics Marx stressed over and over again that the industrial capitalists deliberately abstained from taking the political power because they feared that the concentration of economic and political power in one class would sharpen the class struggle and lead to a social revolution. The short-term economic losses incurred by leaving power in the hands of the aristocracy were seen as smaller than the political gains (and the long-term economic gain) achieved by diverting attention and unrest away from the capitalist class and blurring the lines of class conflict [82].

This explanation does not count as an analysis of precommitment, because neither criterion (*i*) nor criterion (*v*) of Section II above are fulfilled. We can bring out this point by staging a confrontation with some other theories of English politics and Marx's own account of French politics. In the *Economist* there appeared in 1862 an article with the striking title "The advantage to a commercial country of a non-commercial Government" [83], where the author (Walter Bagehot himself?) argues forcefully that the commercial interests should abstain from taking the political power if they know their own good, because the aristocracy is much more capable of taking a disinterested and long-term view than the "rule of wealth" [84]. Similar views were voiced by Hyppolite Taine [85]. The same analysis was offered by Schumpeter when he concluded that the rule of the aristocracy was beneficial for the bourgeoisie because it "needs a master" [86] to protect it against itself. In these accounts it is fairly clear that criterion (*i*) of Section II is fulfilled. They all assume that the bourgeoisie deliberately restricted (or refrained from extending) its political power because it could not trust itself to maintain the stable environment required for long-term growth. Some options that might otherwise have been feasible, were excluded by this strategy. No such restrictions are implied by Marx's theory: the bourgeoisie abstains from power in order to manipulate the working class and not in order to control itself through a self-imposed master.

It seems quite clear that neither Marx's view of English politics nor the alternative theories I have mentioned satisfy criterion (*v*) of Section II. It is absurd to say that the industrial capitalists abdicated from the power: they had never had it. One cannot even argue with much plausibility that there was a deliberate decision not to take power, even though the fact that the Anti-Corn-Law League did not convert itself into a political party could be interpreted as a decision of this kind [87]. It seems more probable that the best conceptualization would be to say that the aristocracy was free to maximize its own objectives with a reasonable profit to the capitalist class as a side constraint. This implies that the capitalist class *would have taken* the political power if its interests had been seriously endangered, but this counterfactual formula is much less substantial than the hard evidence needed for satisfaction of criterion (*v*).

When we go on to Marx's analysis of French politics, and especially of the *coup d'État* of Louis Bonaparte, we see that it has much in common with Schumpeter's view of English politics, with the added implication that an act of abdication can be empirically proved to have taken place. The French bourgeoisie needed a master; they found him in Louis Bonaparte; they abdicated from power and transferred it to him:

> "Thus, by now stigmatizing as '*socialistic*' what it had previously extolled as '*liberal*', the bourgeoisie confesses that its own interests dictate that it

> should be delivered from the danger of its *own rule;* that in order to restore tranquillity in the country its bourgeois Parliament must, first of all, be given its quietus; that in order to preserve its social power intact its political power must be broken; that the individual bourgeois can continue to exploit the other classes and to enjoy undisturbed property, family, religion, and order only on condition that his class be condemned along with the other classes to like political nullity; that in order to save its purse it must forfeit the crown, and the sword that is to safeguard it must at the same time be hung over its own head as a sword of Damocles." [88]

Surely it is only an accident that this magnificently rhetoric passage does not add the metaphor of Ulysses and the Sirens to the others. It *is* rhetoric, however, rather than analysis. Even if we concede Marx the (by no means uncontroversial) point that the anti-capitalist nature of the Bonapartist regime was beneficial or even indispensable for the survival of capitalism as an economic system, it still remains to be shown that this *explains* the emergence and maintenance of that regime. Abstractly speaking two such explanations are possible. Either we adopt the functional mode of analysis and explain the cause (the regime) by its effects; I shall not here dwell upon what I see as the weakness of this line of argument [89]. Or we adopt the intentional mode of analysis and explain the Bonapartist regime as being deliberately set up by an act of abdication from power by the capitalist class. In this case, however, empirical evidence for such deliberations is required; as far as I know none has been forthcoming. The reason why it might appear in this case as if criterion *(v)* is fulfilled is the undeniable discontinuity of the *coup d'État*. The continued political supremacy of the English aristocracy could at most be seen as an act of omission of the English bourgeoisie, but the abrupt change of power from a liberal regime to a Bonapartist one satisfies at least one condition for this to be an act of commission of the French bourgeoisie. It is, however, at least as plausible to see the real actor in Louis Bonaparte himself: to say that the weakness of the bourgeoisie permitted his rise to power is not to say that the bourgeoisie permitted him to take power. To abdicate immediately before you are dethroned may improve appearances, but does not change the substance.

We should conclude, then, that the analysis of democracy has offered some convincing examples of political precommitment; the analysis of the capitalist state has not. We should stress once more the crucial notion of the constituent assembly; not only as a fictional device (as in contract theories of the state) but as a real historical assembly that deliberately seeks to bind its successors. We do not know (or at least I do not know) how the Greek instituted their *graphe paranomon;* whether it was the decision of some *Ur*-assembly or something that slowly crystallized into a custom without having ever been the subject of deliberation. We do know something, however, about the constituent assemblies of the last two hundred years; they are the closest analogy in society to the state of mind of Ulysses before setting out on that dramatic part of his journey.

IX

We are now in a position to draw some conclusions and to pose some further questions. Let us first review the gamut of strategies of precommitment that have been touched upon in the course of this essay:

1. Manipulation of the feasible set
   a) Restricting the set of physically feasible actions
   b) Changing the reward structure by public side bets.
2. Manipulation of the preference structure and of the character
   a) Strengthening the willpower
   b) Changing the structure of desires
3. Manipulation of information
   a) Changing the belief system
   b) Avoiding exposure to certain signals

To this we should add a strategy that fails to satisfy criterion *(iii)* of Section II above, but that nevertheless is a response to the same problem of the weakness of will:

4. Manipulation through a rearrangement of inner space
   a) Using private side bets
   b) Using consistent planning

Let us briefly comment upon these strategies and substrategies. Within the set of actions that change the feasible set we should distinguish between, say, the strategy of going for a walk in the mountain so as to make cigarettes physically unavailable and the strategy of telling your friends that you will stop smoking so as to change the reward system. The latter strategy also induces a change in the feasible set, because the option "Continue to smoke without any sarcastic comments" now becomes unavailable. Within the set of actions that change your character we should distinguish between the general strategy of strengthening the willpower and the more specific strategy of modifying some particular desire: the first permits you to climb higher uphill slopes while the second reduces the height of the slope that is to be climbed. Within the set of actions that modify the information upon which further decisions are taken, we should distinguish between the very radical strategy of inducing new factual beliefs (which includes a strategy for inducing forgetfulness about the induction) and the very moderate strategy of screening yourself from certain signals or cues in the environment. It is *not* a matter of avoiding exposure to information in the strict sense, which is almost as paradoxical as the induction of new beliefs[90]; only a question of avoiding exposure to conditions that might trigger off the efficacy of information which you already possess. You cannot deliberately avoid tobacco shops for the purpose of inducing forgetfulness about tobacco; on the contrary the need for constantly being on the outlook for such shops in order to cross the street before you come to them will more probably tend to strengthen your awareness of tobacco. Nevertheless exposure to tobacco in the concrete might be more tempting than the awareness of the tobacco in the abstract, so that you might accept an increase of the latter for the sake of reducing the former.

The most important of these strategies of precommitment would seem to be methods (1b) and (2b). In most important cases strategy (1a) is unfeasible or involves too heavy sacrifices, so that in most cases it would fail to satisfy criterion *(iv)*. The same, I think, often holds for strategy (2a), which in most cases would be a form of overkill whose feasibility as a solution would do away with the problem itself. (Cp. also the remarks on strategy (4b) in Section V above.) We should note, however, Ainslie's profound suggestion that strategies (4a) and (2a) may be very closely related, *i.e.* that the technique of private side bets may involve a general strengthening of the willpower. When the subject rearranges his reward system so that a failure to follow up one decision (*e.g.* the decision to stop smoking) has negative implications for other decisions (*e.g.* the decision to diet), the interrelated system of goals and desires may come to exhibit the rigidity and inflexibility that are often associated with the notion of willpower[91]. As for strategy (3a), it would seem to be so costly as to be chosen only in such extreme cases as discussed in Section III above; strategy (3b) is frequently used but of doubtful efficacy[92].

In many and perhaps the most important cases the strategy involved implies a combination of methods (1b) and (2b). Realizing that through a series of actions a_1, a_2... a_n I can achieve a *hexis* from which the desired actions will flow naturally ("sans violence, sans art, sans argument"), and that each of these actions is not within immediate reach of my willpower, I may precommit myself to them by changing the reward system. This is how most people go about stopping to smoke. To bring about the state of non-addiction simply by abstaining from smoking may be too hard; to set up a permanent system of public side bets too complicated, and in any case unnecessary if a temporary system can bring about that state and thus make itself superfluous.

"After the age of forty one is responsible for one's face." This truth, equally hard to confront and to dispute, is but one aspect of the general fact of *responsibility for self*. One can never say "This is simply how I am" as an excuse for conduct, for one could have made oneself into a different kind of person. In the philosophical litterature this responsibility has been asserted from two different points of view. One, ultimately of Aristotelian origin, places the self squarely in the world of causal processes and argues that it can and should shape itself by exploiting these processes. This is the approach of the present essay. Another, represented by Sartre and more recently by

Charles Taylor[93], assumes that you can change yourself by redefining and redescribing the self. This would seem to be closer to the techniques of private side bets and of consistent planning. As so many other writers Taylor illustrates his approach by the problem of dieting:

> "Let us take the case of the man who is fighting obesity and who is talked into seeing it as a merely quantitative question of more satisfaction, rather than as a matter of dignity and degradation. As a result of this change, his inner struggle itself becomes transformed, it is now a quite different experience. The opposed motivations — the craving for cream cake and his dissatisfaction with himself at such indulgence — which are the 'objects' undergoing redescription here, are not independent [...] When he comes to accept the new interpretation of his desire to control himself, this desire itself has altered. True, it may be said on one level to have the same goal, that he stop eating cream cake, but since it is no longer understood as a seeking for dignity and self-respect it has become quite a different kind of motivation [...]
>
> Thus our descriptions of our motivations, and our attempts to formulate what we hold important, are not simply descriptions, in that their objects are not fully independent. And yet they are not simply arbitrary either, such that anything goes. There are more or less adequate, more or less truthful, more self-clairvoyant or self-deluding interpretations"[94].

Within constraints, saying makes it so: this is the important truth expressed here. In the social sciences this has been studied under the heading of self-fulfilling predictions, and Brouwer's fixed-point theorem has been invoked to show that in at least some important cases it is possible to describe the situation in a manner that takes account of the fact that the description modifies the situation being described[95]. The crucial point of Taylor's passage, to which nothing corresponds in these studies, is that *there may be several fixed points;* several mutually incompatible ways of describing the situation each of which becomes true by virtue of being asserted. There are constraints that prevent any description from being a self-fulfilling one; these would presumably correspond to the higher-order character traits, and as such must also be subject to the thesis that saying makes it so (within further constraints, and so on). Asserting "This is how I am" is part description, part discovery, and part creation.

The Aristotelian and the Sartrian approaches supplement each other rather than contradict each other (at least if we attribute more importance to the constraints than Sartre seems willing to do). The strategy of precommitment can be used to realize a state that is not attainable through redefinition of the situation. This holds at all levels: using strategies of precommitment it is possible to change the constraints that define the set of possible self-fulfilling descriptions, over and above the extent to which these constraints can themselves be modified by redescription. I, for one, believe that the constraints limiting the self-fulfilling descriptions are very strong, and become increasingly so when we go from the first-order descriptions of the character traits to the higher-order descriptions of the first-order constraints. This implies a broad scope for strategies of precommitment. Anyone, however, that admits that there are some limits to the principle that saying can make it so, must give *some* scope to the devices for binding oneself. The picture that can be seen both as a duck and as a rabbit, but not as both simultaneously, can serve as an analogy here [96]. The lines in the picture serve as a constraint for how the picture can be seen, but adding an extra line can change this constraint in what may be a drastic manner. Here the "act" of switching from duck-vision to rabbit-vision corresponds to a redefinition of the situation, and the act of adding an extra line corresponds to the strategy of precommitment. Responsibility for self includes both types of action.

Before concluding I would like to mention two disturbing problems to which I see no easy solutions. The first can be summed up in a phrase from William James: "the *highest* ethical life [...] consists at all times in the breaking of rules which have grown too narrow for the actual case" [97]. The lowest form of ethical life would presumably be total and myopic impulsiveness; at a higher level is the life according to self-imposed rules or strategies of precommitment; and at the very highest level the deliberate breach of these rules when, all things considered, this seems justified. In order to see clearly where the problem lies in such cases, we may distinguish between at least five senses of "action not governed by rules":

1. Actions performed by a person who lacks the concept of a rule.
2. Actions performed by a person who typically acts according to rules, but who either
   (a) fails to follow the rule in a particular case; or
   (b) decides to break the rule in a particular case
3. Actions performed by a person who has decided never to act according to rules, and who either
   (a) has never followed the rules; or
   (b) has followed the rules and decided to abandon them.

To see the relevance of these distinctions, consider the problem of free vs. bound verse. Most poets writing bound verse would argue, I think, that this is a necessary limitation, and that the boundless freedom of free verse is destructive rather than creative. Only by restricting the infinitely large set of possible statements to the more manageable (but still very large) subset of statements satisfying the formal requirements of rhyme and meter can the poet create the setting where he can exercise his gift of choosing between such statements. The strict form of the sonnet permits the poet to "put Chaos into fourteen lines and keep him there" — and a similar notion no doubt lies behind the feat performed by George Perec in writing a whole novel where the letter "e" is nowhere used.

Against this classical line many other stances are possible. Some would argue that in order to write free verse one should have mastered the technique of bound verse (case 3b), and they would greet with scepticism the brand of poet that claimed to be able to dispense with this mastery (case 3a). Writers of both these persuasions would agree, however, that poetry is essentially different from mere talk (case 1) and that rules — if rules there be — should be mastered completely (unlike case 2a, corresponding to weakness of will). The most complex position would be that the highest poetical life is achieved by bound verse broken by the occasional irregularity when it is required. It is important to bind oneself, to escape from the abyss of possibilities unlimited, but equally important to have a feeling for the occasions when perfection demands liberty of movement unfettered by rules, as in case 2b. This last attitude would seem to capture the idea expressed by William James, though case 3b might also be a possible interpretation.

In the ethical, as distinct from the esthetic case, the main problem is how to distinguish between cases 2a and 2b; how to distinguish between breaking the rules for good reasons and breaking them for bad reasons. How can we know — or how can other people know — if a given piece of impulsive (not rule-governed) behaviour belongs to the highest or to the lowest level? Exceptions recognized in advance as exceptions present no problems; what we want is a criterion for what *would have been* recognized in advance as a legitimate exception if the issue had been raised. Ainslie suggests [98] that in such cases we need a *bright line* (a focal point) in our inner space to tell us if a given exception is an ad-hoc rationalization or a genuine one. Internalization of parental norms might be important here to achieve this ideal of *control without rigidity*.

This problem can also be seen from the point of view of the persons that have the task of binding us: under what conditions should they release us from the mast when we beg them to? D. Parfit has constructed the following example:

> "Let us take a nineteenth-century Russian who, in several years, should inherit vast estates. Because he has socialist ideals, he intends, now, to give the land to the peasants. But he knows that in time his ideals may fade. To guard against this possibility he does two things. He first signs a legal document, which will automatically give away the land, and which can only be revoked with his wife's consent. He then says to his wife, 'If I ever change my mind, and ask you to revoke the document, promise me that you will not consent.' He might add, 'I regard my ideals as essential to me. If I lose these ideals, I want you to think that *I* cease to exist. I want you to regard your husband, then, not as me, the man who asks you for this promise, but only as his later self. Promise me that you would not do what he asks." [99]

Parfit then goes on to argue that according to one plausible way of thinking, "she can never be released from her commitment. For the self to whom she is committed would, in trying to release her, cease to exist"[100]. This ingenious variation on the theme of *Catch 22*[101] does not, perhaps, carry total convincing power. It raises the same ethical questions as the problem whether we are obliged to keep deathbed promises. While I can see the rule-utilitarian arguments for this obligation, I also feel that the very act of exacting such promises may be morally doubtful. These, however, are difficult matters, and not really germane to our problem. Instead of constructing a case of several successive selves, as in Parfit's example, we should imagine a situation where several selves coexist simultaneously and hierarchically, representing the lowest, the intermediate and the highest forms of ethical life as defined above. The question is then how other persons can decide which is the authorized spokesman for the hierarchy; how they can decide whether the revocation of an order is issued by the lowest or the highest self. Other individuals have no direct access to *our* bright lines. Parfit seems to think that this does not pose any problem; that one should never give in to a person begging me to give him the cigarettes he has begged me to withhold from him[102]. This, however, is to beg the very question raised by Ainslie and by James, which is that sometimes I may have good reasons for asking to be released, in the sense of reasons that I would have accepted before I asked to be bound.

The second problem is even more murky. It concerns the spectre of infinite regress that inevitably must arise as soon as we introduce the notion of self-manipulation. We have discussed numerous cases of three-tiered selves, and in principle there seems to be no reason why this could not be extended to any number of levels. It is true that the number three crops up very often, and that plausible four-level cases are hard to construct. (Freud's trinity of id, ego and supergo would correspond roughly to the lowest, highest and intermediate forms of ethical life respectively[103].) Nevertheless the logical possibility of a hierarchy with indefinitely many levels is a disturbing one, and it is hard to see what one should say to a person begging not to be released in the following terms. "The rule-governed self that now speaks to you is one level above (and not one level below) that splendidly arrogant self who, invoking the Jamesian notion of the highest ethical life, begged you to release him. I grant you that this disrespect for rules is justified in the case of my neurotic superego, but not in the case of the moral rules that I adopt on the grounds of the categorical imperative and in the light of pure reason." And going from the point of view of the observer to the point of view of the actor, can I *really* be sure that James' notion is not a temptation which I should resist? How can I know that my bright line is not constantly displacing itself "behind my back" so as to justify *ex post* any exception to the rule?

I have no idea that would count as a solution to this problem, but I have a hunch about the lines along which a solution is most likely to be found. We should invoke, I think, an analogy from game theory, where the apparently infinite regress of "I think that he thinks that I think that he thinks..." does not prevent a game from having a unique and predictable solution, even in the absence of dominant strategies[104]. This approach is advocated, for example, by H.G. Frankfurt in a well-known contribution to the debate[105]. I am not sure that I understand his argument for (or indeed his version of) this solution, but I believe that this general kind of analysis is the only one that offers some solid hope of escape. Perhaps one could appeal to fixed-point reasoning, in order to prove that for every person (or for some persons only?) there exists a level beyond which the preference structure repeats itself identically at every level. I am fully aware that this suggestion is both extremely obscure and (to the extent that it is intelligible) pregnant with new problems, but I propose to leave the question at this point.

For a conclusion I think it suffices to repeat a point made in the discussion of criterion *(iv)* in Section II above. A full characterization of what it means to be human should include at least three features. Man can be *rational*, in the sense of sacrificing present gratification for future gratification. Man often is not rational, and rather exhibits *weakness of will.* Even when not rational, man knows that he is irrational and is able to *bind himself* to protect himself against the irrationality. The second-best or imperfect rationality takes care both of reason and of passion. What is lost, perhaps, is the sense of adventure.

| *Jon Elster is Lecturer at the University of Oslo, Norway.*

Notes

[*] I am grateful for the comments by Amélie Rorty, Arthur Stinchcombe and George Ainslie on an earlier version of this paper.

[1] R.H. Strotz, "Myopia and inconsistency in dynamic utility maximization", *Review of economic studies*, 1955-56.

[2] G. Ainslie, "Specious reward: A behavioral theory of impulsiveness and impulse control", *Psychological bulletin*, 1975. The notion of private side bets proposed in this paper is further explored in Ainslie's "A behavioral understanding of the defence mechanism", mimeo, 1977.

[3] Stendhal, *Romans et nouvelles*, Paris, Gallimard, La Pléiade, vol. 1, 1952, p. 969.

[4] *Ibid.*

[5] T. Nagel, *The possibility of altruism*, Oxford, Oxford University Press, 1970, p. 73.

[6] Ainslie, "Specious reward", p. 478. One of his examples goes as follows. "Another example might be a person with antisocial impulses who has found that acts of delinquency put him into unrewarding situations and thus tries to avoid the temptation to perform them. However, if he is afraid that he may get out of control and get himself into bigger trouble, he may stop avoiding the temptation to act up in smaller ways, so that the authorities (police,

ward attendants, etc.) will exert more effort to guard him. The long-range desire to avoid a major rampage causes him to look for devices that will constrain his future behavior; because such a device must be chosen early, when the effectiveness of the reward it leads to is low, one that also produces an immediate thrill may be choosable, while one that stands on its own may not be."

[7] Wittgenstein, *Philosophical investigations*, Oxford, Blackwell, 1953. Paragraph 265.

[8] Ainslie, "Specious reward", p. 478 ff. and "A behavioral understanding", p. 18 ff. and *passim*. A brief explanation of this concept may be of interest. We assume that present effectiveness of future reward is a decreasing function of the time from now till the reward is due. We assume, furthermore, that the curve relating time to effectiveness is more concave than an exponential curve, *e.g.* that it is hyperbolic in form. (See Section V for the privileged nature of exponential curves.) This means that the relative effectiveness of a large reward in the distant future and a smaller reward in the near future may change when the corresponding curves cross (which they can never do if they are exponential). There is a switch point, that is, when the person stops preferring the large reward and begins preferring the smaller reward. Ainslie then imagines a case where a subject is offered a choice between two such rewards a number of times in succession: the private side bet is then a decision to choose between all pairs simultaneously instead of making a succession of choices. It can then be shown that this method of decision delays the switch point and increases the probability that the larger reward will be chosen. If I can persuade myself to think that *either* I shall always resist the temptation to take a second helping of dessert *or* I shall never do so, then the chances are better that I shall never do so than if I make a separate choice at each occasion. Cp. also note 26 below, where the notion of *public* side bet is briefly defined.

[9] "Beyond gradient-climbing", paper presented to the Fourth International Congress of the International Organization for the Study of Human Development, Paris, 1977. A French version ("Critique des analogies socio-biologiques") will appear in *Revue française de sociologie*, 1977.

[10] C. Azzi and R. Ehrenberg, "Household allocation of time and church attendance", *Journal of political economy*, 1975. For another reason that could be invoked to prove that eternal bliss adds up to a finite total, see my *Leibniz et la formation de l'esprit capitaliste*, Paris, Aubier-Montaigne, 1975, p. 148.

[11] *Pensée* 233.

[12] Cp. Section IX below for a brief discussion of this problem. The point is that in some cases the belief may not only change the world in some (desired) manner but change it in a manner that makes the belief come out true. More formally, let us assume that the state of the world y is a function f of my beliefs about the world x (and of a number of other factors that are kept constant for the purposes of this analysis). Let us further assume that my utility u is a function g of y. Let us also assume that the conditions of Brouwer's fixed-point theorem are fulfilled, so that there exists an x such that $x = f(x)$. We then define $\bar{x}$ as the value of x that maximizes u = g (f (x)). Adopting a self-fulfilling belief for its causal efficacy would then require $x = \bar{x}$, which could only arrive by accident. There is also a second possibility, viz. in the case where there are *several* fixed points $x_1, x_2 \ldots x_n$. (This' of course, would also be a mere accident, as Brouwer's theorem only states the existence of at least one fixed point.) Then the utility function u could be invoked in order to choose between these points, so that one should choose the x_i giving the largest value u (x_i) even if this would typically be a lower value than u ($\bar{x}$). Cp. also note 95 below.

[13] A. Stinchcombe, *Constructing social theories*, New York, Harcourt, Brace and World, 1968, p. 116.

[14] From *Le rouge et le noir*, in: Stendhal, *Romans et nouvelles*, vol. 1, pp. 317-318.

[15] "Deciding to believe", in: B.A.O. Williams, *Problems of the self*, Cambridge,

Cambridge University Press, 1973. Cp. also D. Pears, "Freud, Sartre and self-deception" in: R. Wollheim (ed.), *Freud: A collection of critical essays*, New York, Anchor Books, 1974, especially p. 105 ff. Pears points to a problem that, when applied to Pascal, might be formulated in the following manner: in the gradual process of a growing belief and a dwindling reason, might there not come a point where the first is not yet strong enough to support the religious behaviour and the second is no longer strong enough to do so? A rather different case where it might seem rational or desirable to manipulate yourself into believing something for which you lack rational grounds of belief, is the so-called "Newcomb's problem"; cp. my discussion of this issue in my *Logic and society*. London, Wiley, forthcoming, chap. 4.

16. E. Dickinson, *Complete poems*, nr. 1560. Williams ("Deciding to believe", pp. 150-151) sees an asymmetry between deciding to forget and deciding to believe, but actually I think his argument confuses the decision to forget something which I have known and the decision not to acquire some new knowledge which I feel I can manage without (such as the number of inhabitants in some distant country). There is no reason why I should attempt *complete* knowledge, but there are good reasons for a *cumulative* acquisition of knowledge. It is true that there are many acquired pieces of knowledge that are so unimportant that it does not matter if they are lost, but by hypothesis this is not true of a piece of information that is considered worthy of being deliberately forgotten. This is the point of Emily Dickinson's poem.

17. An example from double-bind theory might be the mother saying to her daughter, "Remember that you are not even to think of this forbidden thing", which is an injunction that the forbidden thing shall both be forgotten and kept in mind. The injunction belongs to the same class of statements as "Don't be so obedient" or "Why do you never bring me flowers except when I ask you to?"—trying to bring about deliberately what is by definition a spontaneous action. (P. Watzlawick, *How real is real*, New York, Vintage Books, 1977, chap. 2. Cp. also note 101 below.)

18. *Pensée* 252.

19. Cp. J. Needham, *Science and civilisation in China*, Cambridge, Cambridge University Press, vol. 2, 1956, p. 16 ff., p. 43 ff. and vol. 4 (3), 1971, p. 227 ff. for the use of these analogies in Chinese moral philosophy.

20. *The Nicomachean ethics*, 1147.

21. For discussions of this conceptual problem see G. Mortimore (ed.), *Weakness of will*, London, Macmillan, 1971 and (especially) D. Davidson, "Weakness of the will" in: J. Feinberg (ed.), *Moral concepts*, Oxford, Oxford University Press, 1969. Davidson concludes that weakness of will is a form of *surdity* where the causal processes of the mind operate behind the back, as it were, of the conscious mind. The reasons which *cause* me to do x may prevail over the reasons that are *reasons* for doing y, even when the latter are stronger (qua reasons) than the former. I do not know how the notion of imperfect rationality should be expressed in Davidson's framework (which seems to be the best yet proposed), but I do not think there is any incompability between the two approaches.

22. *Pensée* 272.

23. Descartes, *Œuvres*, ed. Adam et Tannery, vol. 4, p. 357.

24. *Les passions de l'âme*, paragraph 48.

25. *The passions and the interests*, Princeton, NJ, Princeton University Press, 1977, especially p. 20 ff.

26. A public side bet is an irreversible and irrevocable change in the payoff structure, in contra-distinction to the private side bets that may be reversed or revoked at will (note 8 above). In bargaining games where there is no well-defined non-cooperative solution (such

as "Chicken") the method of public side bets can be used to great effect. Consider, for example, the two games below:

Chicken

		Actor 2	
		Swerve	Do not swerve
Actor 1	Swerve	3,3	1,5
	Do not swerve	5,1	0,0

Chicken with side bets

		Actor 2	
		Swerve	Do not swerve
Actor 1	Swerve	1,3	-1,5
	Do not swerve	5,1	0,0

In the second game actor 1 has been able to place a side bet that obliges him to pay 2 units (to some third actor) if he chooses to swerve. This means that his non-swerving strategy is now a dominant one *(i.e.* the best whatever actor 2 decides to do) so that considerations of rationality will compel him not to swerve. Knowing this, actor 2 will be compelled by the same considerations to swerve. This means—paradoxically—that by worsening some of the outcomes actor 1 ensures for himself a better outcome than he could have hoped for otherwise. There are, of course, two snags. In the first place the credibility of the side bet is all-important: it must be entrusted with some impartial agency which could be counted on to collect the fine whatever happens. In the second place you must place your bet (and make it known) before your opponent gets the same idea. If both actors place side bets of 2 simultaneously, the game becomes a Prisoner's Dilemma, where both are compelled to the head-on course leading to mutual destruction. It is a surprising fact, worthy of explanation, that no society (to my knowledge) has developed institutions (private or public) specializing in accepting side bets for the purpose of accomodating people's wishes to bind themselves.

27. *Les passions de l'âme*, paragraph 50. Italics added.

28. "Only an utterly senseless person can fail to know that our characters are the result of our conduct; but if a man knowingly acts in a way that will result in his becoming unjust, he must be said to be voluntarily unjust." *(The Nicomachean ethics*, 1114). This observation seems correct in the standard case where the character *(hexis)* set up by a series of actions just is the disposition to perform that kind of action. We certainly should know that smoking will bring about addiction to smoking. The problem, however, is that the working of the mind may also be subject to more complex laws. Repetition of some action may bring about a disposition *not* to perform that action, and this may be exploited for purposes of self-control as when people saturate themselves with sweets in order to bringabout an aversion to sweets. It may be difficult to know in advance which acts of repetition are habit-forming and which are habit-curing. I am certain that there are other mechanisms even more complex, as when a series of actions A interact with an existing disposition B so as to produce a tendency o perform actions C. (Does use of marijuana lead to heroin?) It would not only be unrealistic to require that people have complete knowledge about all such mechanisms; the requirement would be logically inconsistent, because the causal processes involved are themselves (higher-order) dispositions that can be modified through conduct, and so on. To this we should add the elusive character of some character traits, that disappear once understood (innocence is the standard example, but there are many others). The injunction "Know thyself" harbours even more paradoxes, which are partly and briefly discussed in Section IX below.

29. Descartes, *Œuvres*, vol. 6, p. 24.

30. See for example W.H. Riker and P.C. Ordeshook, *An introduction to positive political theory*, Englewood Cliffs, NJ, Prentice-Hall, 1973, chap. 2.

31. H. Simon, "A behavioural theory of rational choice", *Quarterly journal of economics*, 1954.

32. *Leibniz et la formation de l'esprit capitaliste*, chap. 3.

33. D. North, "Institutional change and economic growth", *Journal of economic history*, 1971, p. 122. A spirited defense of this approach is to be found in G. Becker, *The economic approach to human behavior*, Chicago, Ill., Chicago University Press, 1976, especially chap. 1.

34. Cp. note 9 above.

35. S.N. Winter, "Economic 'natural selection' and the theory of the firm", *Yale economic essays*, 1964, p. 262. Cp. also by the same writer, "Optimization and evolution", in: R.H. Day and T. Groves (eds.), *Adaptive economic models*, New York, Academic Press, 1975.

36. S.M. Robbins and R.B. Stobaugh, *Money in the multinational enterprise*, London, Longman, 1974, p. 130.

37. C. Tisdell, "Economic policy, forecasting and flexibility", *Weltwirtschaftliches Archiv*, 1971, p. 34-35.

38. "On the optimal use of forecasts in economic policy", *Journal of public economics*, 1972.

39. W. Nordhaus, "Tne political business cycle", *Review of economic studies*, 1975, p. 188. The problems raised in the last two sentences of this passage are discussed in Section IX below, under the heading of what William James called "the highest ethical life".

40. A. Lindbeck, "Stabilization policy in open economies with endogenous politicians", *American economic review*,—Papers and proceedings, 1976, p. 18, note 8.

41. A.O. Hirschman, *The passions and the interests*, p. 50.

42. Descartes, *Œuvres*, vol. 11, p. 34.

43. *Ibid.*, vol. 4, p. 316.

44. *Ibid.*, vol. 4, pp. 356-57.

45. G. Becker, *The economic approach to human behavior*, chaps. 12 and 13. See for example p. 288: "an egoist has an incentive to try to simulate altruism whenever altruistic behavior increases his own consumption through its effect on the behavior of others".

46. See especially M. Taylor, *Anarchy and cooperation*, London, Wiley, 1976.

47. P. Hammond, "Changing tastes and coherent dynamic choice", *Review of economic studies*, 1976. He argues (on p. 163 of this paper) that precommitment "is not really a way of resolving inconsistency", and that it "means no more than determining what really *is* feasible at any stage". He then retracts (and for our purposes to a sufficient extent) by adding that "such an exercise is often far from trivial; indeed, it is often like a kind of technical progress". I have no quarrel with this characterization: consistent planning is sophistication within the limits of the feasible, precommitment is sophistication amounting to a modification of the limits.

48. This is strongly argued by J. Rawls, *A theory of justice*, Cambridge, Mass., Harvard University Press, 1971, p. 284 ff. for the inter-generational case (but see next footnote) and by T. Nagel, *The possibility of altruism* for the intra-generational case. Ingenious arguments for the contrary view have been proposed by D. Parfit ("Later selves and moral principles", in: A. Montefiore (ed.), *Philosophy and personal relations*, London, Routledge, 1973) and by B.A.O. Williams ("Persons, character and morality", in: A. Rorty (ed.), *The identity of persons*, Berkeley and Los Angeles, University of California Press, 1976). For the sake of argument I here assume the Rawls-Nagel view, so that time preferences always constitute a *problem*. I should add, however, that there is much to be said for the Parfit-Williams view, but presumably most of my arguments would be accepted even by them for the cases (and

surely there are such cases) when time preferences correspond to weakness of will and not (as in the cases discussed by Williams) to character traits without which my life is without substance.

49. If a society, for example, tried to choose the saving rate that maximizes total (undiscounted) consumption over infinite time, it would soon find out that there is no such saving rate: for any rate below 100% there is a higher rate that gives larger total consumption, but the 100 % rate itself gives zero consumption. (This statement, of course, depends on a number of assumptions that the reader will find spelled out in any textbook on planning, *e.g.* G.M. Heal, *The theory of economic planning*, Amsterdam, North-Holland, 1973.)

50. E.S. Phelps and R. Pollak, "On second-best national saving and game-theoretic equilibrium growth", *Review of economic studies*, 1968.

51. *Confessiones*, VIII, *vii*.

52. R. Pollak, "Consistent planning", *Review of economic studies*, 1968.

53. As observed by G. Becker (*The economic approach to human behavior*, chap. 8) there are at least two alternatives to the rational-choice mechanism for the second filtering process: traditional behaviour and random behaviour. A logical possibility might be, therefore, that an individual tried to precommit himself by changing the mechanism from rational choice into one of these two alternatives. Abstracting from this (very abstract) possibility, I shall assume that the change is always from rational choice according to one set of preferences into rational choice according to a different set of preferences.

54. The most recent and thorough discussion is found in the contributions to the Symposium on "Formed habits", *Journal of economic theory*, 1976. Earlier contributions include T. Haavelmo, "The probability approach in econometrics", *Econometrica*, 1944 (supplement), p. 17 ff.; N. Georgescu-Roegen, "The theory of choice and the constancy of economic laws", *Quarterly journal of economics*, 1950; W.M. Gorman, "Tastes, habits and choices", *International economic review*, 1967; M.H. Peston, "Changing utility functions", in: M. Shubik (ed.), *Essays in mathematical economics, in honor of Oskar Morgenstern*, Princeton, NJ, Princeton University Press, 1967; R. Pollak, "Habit formation and dynamic demand functions", *Journal of political economy*, 1970; C.C. von Weiszäcker, "Notes on endogenous change of tastes", *Journal of economic theory*, 1971; R.M. Cyert and M.H. de Groot, "Adaptive utility", in: R.H. Day and T. Groves (eds.), *Adaptive economic models*. A sceptical note is sounded by G. Becker and G. Stigler ("De gustibus non est disputandum", *American economic review*, 1977) who argue that such phenomena as habit-formation and addiction are quite compatible with stable preferences. A person addicting himself to good music does not change his preferences, only his consumption technology by accumulating an invisible capital that enables him to derive more enjoyment from a given amount of time spent listening to music.

55. See preceding note.

56. R. Pollak, "Habit formation and long-run utility functions", *Journal of economic theory*, 1976.

57. This point is also made by Pollak, *ibid.*

58. Ainslie, "Specious reward", pp. 472-473.

59. *Ibid*, p. 476.

60. See the reference in note 9 above.

61. E.O. Wilson, *Sociobiology*, Cambridge, Mass., Harvard University Press, 1975, p. 171.

62. M.I. Finley, *Democracy: Ancient and modern*. London, Chatto and Windus, 1973, p. 26.

63. *Ibid.*, p. 80.

64. *Ibid.*, p. 67.

65. "Myopia and inconsistency", p. 178.

66. For a criticism of functionalist sociology see the reference in note 9 above.

67. F. Sejersted, *Ideal, teori og virkelighet*, Oslo, Cappelen, 1973; cp. also my review of this work (a biography of the Governor of the Bank of Norway between the wars) in *Economic history review*, 1975.

68. This is stressed both by Sejersted *(ibid.)* and by Nordhaus (see the text quoted in Section IV above).

69. Here and in much of the following I draw heavily upon work by Francis Sejersted. In addition to the work cited in note 67 above. I should mention his *Historisk introduksjon til økonomien*, Oslo, Cappelen, 1973 and his "En teori om embedsmannsstaten 1814-1884", Oslo, 1975, mimeo.

70. Needham, *Science and civilisation in China*, vol 2, p. 521 ff.

71. Cp. Leibniz: "il n'importe pas si les lois décident certaines choses touchant le droit des particuliers autrement qu'on ne le déciderait selon la seule raison naturelle. Par exemple, si les abeilles sorties de mon fonds doivent être à celui qui les prend le premier. Et choses semblables. Car le droit étant déterminé, on se peut régler là-dessus. "(Quoted and commented in my *Leibniz et la formation de l'esprit capitaliste*, p. 142.)

72. Thus short-term efficiency provides arguments against the patent system, whereas long-term efficiency (and at least some principles of distributive justive) requires this system.

73. "En teori om embedsmannsstaten", *passim*.

74. Cp. my *Logic and society*, chap. 5 for an extended discussion of this notion.

75. R. Suttmeier, *Research and revolution*, Lexington, Mass., Lexington Books, 1974, p. 91.

76. Sejersted ("En teori om embedsmannsstaten") cites Thomas Jeffersen and the Norwegian Henrik Steenbuch.

77. Cp. W. Baumol, *Welfare economics and the theory of the state*, London, Longmans, 1952; M. Olson, jr., *The logic of collective action*, Cambridge, Mass., Harvard University Press, 1965 and M. Taylor, *Anarchy and cooperation*.

78. In concrete terms, the task of the state should be to facilitate cartels and employers' organizations and to prevent union formation. In appearance the English Anti-combination Acts were directed against employers and workers alike, but in the first place there was a formal asymmetry in the fact that the statutes "only allow of a civil action against the contract-breaking master, but on the contrary permit a criminal action against the contract-breaking workman" (Marx, *Capital*, vol. 1., New York, International Publishers, 1967, p. 740) and in the second place the Acts were enforced to quite different extents against contract-breaking masters and workers.

79. Marx, *Grundrisse der Kritik der politischen Oekonomie*, Berlin, Dietz, 1953, pp. 429-430. Marx does not specify the mechanisms whereby this could come about, but at least two possibilities come to mind. Either some of the firms could achieve a size which would make it profitable for them to provide the public good singlehandedly (cp. M. Olson, *The logic of collective action*, p. 29) or institutional devices (such as the patent system) could be invented to internalize the external economies.

80. *Capital*, vol. 1, pp. 238-39.

81. *The communist manifesto*.

82. See especially K. Marx, "The chartists", *New York Daily Tribune* 25.8.1852 and "Die grosse parlamentarische Debatte", *Neue Oder Zeitung* 12.6.1855.

83. I owe this reference, as well as the other references in this paragraph, to S. Grindheim, "Hvorfor fikk aristokratiet regjere når borgerskapet hersket?", Oslo, 1975 (Master's thesis, mimeo).

84. Quoted by Grindheim from a discussion of the same problem in Bagehot's *The English constitution*, Ithaca, NY, Cornell University Press, 1966, p. 122.

85. Quoted by Grindheim from H. Taine, *Notes on England*, London/The Hague, Thames and Hudson, 1957, pp. 155-156.

86. J. Schumpeter, *Capitalism, socialism and democracy*, London, Allen and Unwin, 1953, p. 139.

87. I owe this suggestion to Kåre Tønnesson.

88. Marx, *The Eighteenth Brumaire of Louis Bonaparte*.

89. In the Marxist brand of functionalism the weaknesses that are common to all varieties of functionalist sociology are aggravated by the arbitrary and inconsistent switch between long-term and short-term effects; cp. the reference cited in note 9 above.

90. For the paradoxes of avoidance of information, cp. my *Logic and society*, chap. 4. Roughly the point is that information about information that disconfirms X *is* information that disconfirms X, so that in order to avoid exposure to negative information you must already have negative information.

91. Ainslie, "A behavioral understanding", explores this at great length.

92. Ainslie, "Specious reward", p. 478.

93. "Responsibility for self", in: A. Rorty (ed.), *The identity of persons*.

94. *Ibid.*, p. 295. The same point is argued in Taylor's "Interpretation and the sciences of man", *Review of metaphysics*, 1971.

95. Cp. note 12 above. For an up-to-date discussion of this problem as it arises in sociology, political science and economics, with full references to the literature, see S. Brams, *Paradoxes in politics*, New York, Free Press, 1976, chap. 3. As far as I know there has been no attempt to link these issues to the analogous problems that arise in psychology. (The Rawlsian notion of a reflective equilibrium could be interpreted in terms of fixed-point reasoning instead in terms of a convergence process, and perhaps some theories of linguistic explanation could also be restated in this way.) I think sociology and psychology could fertilize each other at this point: from sociology the psychologist could get the inducement to state the problem of self-definition in the precise language of Brouwers' fixed-point theorem, which would surely help him to bring out underlying and unstated assumptions; from psychology the sociologist could get the idea that there might be several fixed-points, the choice between which cannot be made on scientific grounds alone.

96. I owe this interpretation of Sartre to Dagfinn Føllesdal.

97. Ainslie, "A behavioral understanding", p. 38 quotes this passage and similar assertions by other writers.

98. Ainslie, *ibid.*, p. 23 and *passim*.

99. D. Parfit, "Later selves and moral principles", p. 145.

100. *Ibid.*, p. 146.

101. Succinctly stated thus by P. Watzlawick: "anybody willing to fly combat missions would have to be crazy, and being crazy, could be grounded for psychiatric reasons. He has only to ask. But *the very process of asking*, of not wanting to fly any more combat missions, is evidence of normalcy and rules out being grounded for psychiatric reasons". (*How real is real*, p. 25, italics added.)

102. Parfit, *op. cit.*, p. 145.

103. I take this suggestion from Ainslie, "A behavioral understanding", p. 38.

104. There are, of course, games where this regress is a real one, such as in the game of Chicken discussed in note 26 above. I am only making the point that in many cases which traditionally were thought to involve this regress, the notion of an equilibrium point and the admission of mixed strategies permit us to short-circuit the regress and converge upon a unique solution.

105. "It is possible, however, to terminate such a series of acts without cutting it off arbi-

trarily. When a person identifies himself *decisively* with one of his first-order desires, this commitment "resounds" throughout the potentially endless array of higher orders". (H.G. Frankfurt, "Freedom of Will and the concept of a person", *Journal of philosophy*, 1971, p.16). I believe that this "resonance effect" should be taken as a *definition* of what it means to commit oneself decisively rather than as being an *effect* of such a commitment. In other words I do not agree with Frankfurt when he goes on to say that "It is relatively unimportant whether we explain this by saying that this commitment implicitly generates an endless series of confirming desires of higher orders, or by saying that the commitment is tantamount to a dissolution of the pointedness of all questions concerning higher orders of desire".