Mutual Advantage Contractarianism and Future Generations

by
GUSTAF ARRHENIUS
Uppsala University

MUTUAL ADVANTAGE CONTRACTARIANS TRY to show that moral principles are best understood as rational limitations on our choices. These limitations are derived from non-moral principles of rational choice. To act rationally would thus be to act morally and vice versa. Consequently, there would be no conflict between individual utility maximisation and morality. The main advantage of this project is that it would answer the moral sceptic's question: Why be moral? The answer: For your own good.

According to the usual conception of rationality, a rational person maximises her utility. There are, however, a number of games that show that if everyone tries to maximise their utility, then everyone will be worse off than they could have been—the only strategic equilibrium in such games are suboptimal. The well-known Prisoner's Dilemma is the paradigmatic example: From the perspective of each individual, it seems rational to squeal. If everyone squeals, however, they will be worse off than they would have been if everyone had kept silent. In response to such problems, contractarians suggest that rational agents should commit themselves to some behaviour that assures them the benefits of cooperation. Principles for such behaviour and principles for the distribution of the cooperative surplus constitute the contractarian moral theory. In the Prisoner's Dilemma, for example, you would commit yourself to not squealing if you have good reason to believe that other prisoners are similarly committed: they will not squeal if you do not squeal.[1]

It has been difficult for mutual advantage contractarians to give a satisfying account of intergenerational justice. The crux of the matter is that the present generation has nothing to gain from a saving policy and future generations cannot wield any threat against previous generations. Thus, contractarianism seems to have the counterintuitive implication that we

[1] This is roughly David Gauthier's solution to the Prisoner's Dilemma. See Gauthier (1986).

ought to adopt depletionary policies whenever it benefits us: We do not have any obligations toward people in the further future.

The foremost mutual advantage contractarian today, David Gauthier, argues that this does not follow. He writes:

> Does not the argument for ignoring those descendants whose lives do not overlap with our own in effect ignore the real significance of generational overlap? The generations of humankind do not march on and off the stage of life in a body, with but one generation on stage at any time. Each person interacts with others both older and younger than himself, and enters thereby into a continuous thread of interaction extending from the most remote human past to the farthest future of our kind. Mutually beneficial co-operation directly involves persons of different but overlapping generations, but this creates indirect co-operative links extending throughout history. Each person, in considering the terms on which he is to co-operate with those in an earlier generation than himself, must keep in mind his need to establish similar terms with those of a later generation, who in turn must keep in mind their need to co-operate with members of a yet later generation, and so on.[2]

An example might shed light on Gauthierís idea: Assume that you and your daughter know that your (prospective) nephew is only going to take care of your daughter at her old age if he has 10 000 pounds in his bank account. Since your daughter realises this, she is not going to spend her money and time on taking care of you in your old age (but rather save it for her own old age) if you are not willing to set aside 10 000 pounds for your nephew. Assuming that it would cost you more than 10 000 pounds to pay for care in your old age, you are going to lose money if you do not set aside 10 000 pounds for your nephew. Consequently, your nephew can, in a sense, punish you if you spend your money rather than saving it for him, even if you and your nephew are not going to be alive at the same time.

Gauthier's attempt to incorporate intergenerational justice in his theory lacks rigour and clarity. Joseph Heath has tried to explicate Gauthier's idea with a repeated Prisoner's Dilemma game.[3] He claims that his model shows that if earlier generations adopt depletionary policies, then the only equilibrium strategy is universal defection which would be worse for everyone. Consequently, it would be irrational for earlier generation to deplete resources at the expense of future generations.

Since Heath's theory involves some technicalities that need not con-

---

[2] Ibid., p. 299.

[3] Heath (1997).

cern us here, I shall present a simpler version which brings out the essential aspects of his reasoning more clearly. I shall then give three examples to show that Heath's model does not rule out depletionary and unfair policies.

The Model

As is usual in mutual advantage contractarianism, people are assumed to be mutually unconcerned maximisers of utility. Cooperation generates a surplus as compared to the state of nature where nobody cooperates. People play an infinitely repeated game where they are able to adopt strategies that make their present actions contingent upon the past behaviour of other players. Generations overlap, that is, people from two or more generations (depending on how one defines the length of a generation) will live at the same time. Assume that people live for eighty years and that they procreate every twenty years (see fig. 1). Let us define a generation as the people born within a twenty year period. With Heath, we assume that the number of deaths are the same as the number of births, that is, that population and generation size are constant. Let us say that during the years 2000–20 there will exist 4 people, p1, p2, p3 and p4 (see column 1). At the end of this period, p1 will reach eighty and consequently die. p2 will reach sixty, p3 will reach forty and p4, who was born in the beginning of this period, will reach twenty. In the beginning of the next period, that is, column 2, p5 will be born and in the end p2 will die. Consequently, a column represents a twenty year period where we have four persons from four different generations present.

Let us represent the utilities of people in different generations by putting the utility in order of ascending index. For example, in column 2, row 2, (7,4,4,7) means that p2's utility is seven during this period, p3's utility is four and so forth.

	1 2000–20 (p1,p2,p3,p4)	2 20–40 (p2,p3,p4,p5)	3 40–60 (p3,p4,p5,p6)	4 60–80 (p4,p5,p6,p7)	5 80–100 (p5,p6,p7,p8)
1. Trigger Strategy	(9,6,6,6)	(9,6,6,**6**)	(9,6,**6**,6)	(9,**6**,6,6)	(**9**,6,6,6)
2. p5 defects in C2	(9,6,6,6)	(7,4,4,**7**)	(3,3,**3**,3)	(3,**3**,3,3)	(**3**,3,3,3)
3. Depletion 1	(3,3,3,3)	(3,3,3,3)	(3,3,3,3)	(3,3,3,3)	(3,3,3,3)

Figure 1

The payoff when x people cooperate is $2x$. The payoff for a defector is $2x+3$. The payoff in the state of nature where nobody cooperates is 3 ($2\times0+3$).

Observe that this matrix is a slice of an iterated game that is presumed to be infinite. If we call the people that play their last game "seniors," then an equilibrium strategy in this game would be what Heath calls a "trigger strategy":

The Trigger Strategy: Always defect in the last game, cooperate in every other game until some non-senior defects, then defect in every subsequent game.

This strategy yields the outcome shown in row 1. This seems to be a reasonable strategy. We can equally well let the people who are playing their last game, the "seniors", defect since they are not taking part in any future game and thus cannot be punished. A non-senior that defects would get a smaller payoff because there would be universal defection in subsequent games (if everybody else has adopted the trigger strategy). For example, if p5 defects in column 2, then his total payoff will be 16 ($7+3+3+3$) as compared to 27 ($6\times3+9$) if he follows the trigger strategy (see row 2).

Let us say that the cooperative venture at stake in these games is a corn field. Let us also add another choice: The people in column 1 could use a fertiliser that would increase the return in that game. In our example, the payoff for cooperation changes to $3x$ in column 1. In subsequent games, the output returns to normal ($2x$) until column 5 when the soil, as result of the use of the fertiliser, is destroyed and cannot produce any corn at all. This is an example of what Heath calls a "depletionary investment policy": A policy "that will, in the foreseeable future, reduce the size of the cooperative surplus ... to less than the state of nature."[4]

What are the equilibrium strategies here? According to Heath, universal defection is the only equilibrium strategy in this scenario, as shown in row 3 (Depletion 1). His argument is based on backward induction: p5 can defect with impunity in column 4 because there cannot be any cooperative surplus in column 5. That is, it will not make any difference for p5 if everybody defects in column 5. The same holds for p6 and p7. We can then repeat this process for columns 3, 2 and 1. The point of Heath's argument is that it would be rational for earlier generations to commit

[4] Ibid., p. 369.

themselves to not adopting depletionary policies since they could then achieve an equilibrium that would give them higher utility. Heath concludes from this that "...*every* generation has an effective threat that it can wield against all previous generations. By simply refusing to cooperate, any generation could ... terminate cooperation in all previous generations. This means that any investment policy adopted must be one that each generation anticipates will be acceptable to all future generations."[5] Moreover, Heath claims that "...there is no problem in securing stable investment across generation under the assumption of mutual unconcern..." and that Gauthier "is perfectly justified in claiming that, within his framework, the only sustainable investment policy is one that is *fair to all future generations*."[6]

Three ways to spend at the expense of future generations

Let us first notice that Heath assumes that the employment of backward induction is unproblematic. I am hesitant, however, to ground a theory of intergenerational justice on such a disputed method as backward induction. Firstly, this method involves extreme demands on the players rationality and their beliefs in their own rationality and other players rationality. The appropriateness of the idealisations involved are undoubtedly problematic. At any rate, Heath has not shown us that the method of backward induction is applicable to his intergenerational game. Secondly, it has been shown that backward induction generates clearly counterintuitive results in a number of cases. In one of the most well known cases—the paradox of the unexpected examination—backward induction assures us that we should not take a professor seriously when she says: "Sometime during next week I shall give you an examination, but in the morning on the day it will occur, you will have no good reason to expect that it will occur that day." The examination cannot take place on Friday (under the assumption that Friday is the last possible day of the week), because when Friday arrives we know that all previous days have been examination free and consequently have every reason to believe that it is going to occur on Friday. In the same vein, the examination cannot take place

[5] Ibid., p. 371 (emphasis in original).
[6] Ibid., p. 374 (emphasis in original).

on Thursday, because on Thursday morning we all know that it has to be held on Thursday because we have already concluded that it cannot occur on Friday, and so forth. In other words, we cannot take the professor's threat seriously. This is, however, false, as all students know. Similarly, the result that Heath gets with backward induction could be considered as another paradox to be used as an argument against backward induction rather than as support for a game theoretical model of intergenerational justice.[7]

Putting the problems with backward induction aside, there are other problems with Heath's theory. It is easy to find examples where Heath's model leaves room for depletionary policies. Heath has only considered depletionary policies of the "doomsday" type, that is, policies that affect all subsequent generations (after the first affected generation) in a manner that makes any future cooperation fruitless (see figure 2).

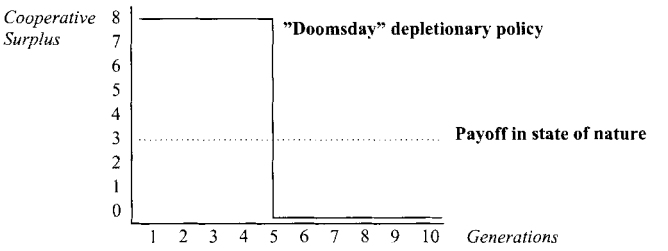

Figure 2

Many possible real world depletionary policies are not of this type. Arguably, depletionary policies could involve negative consequences that only stay in effect for a limited time. If we deplete all oil resources, for example, this could conceivably diminish the cooperative surplus of future generations for a limited time (see figure 3).

Indeed, earlier generations can use depletionary investment policies with impunity as long as the negative consequences of such policies stay in effect for a limited time.[8] Let us take an example where the negative

[7] There is an extensive discussion of the problems with backward induction in the literature. See, for example, Rabinowicz (1996), Sliwinski (1995) and Sobel (1993).

[8] One can show that compliance of future generations can be secured if the negative effects stay in effect for no more than $g-2$ generations, where g is the number of generations that exist at the same time. In our simplified model, where four generations live at the same time, deple-

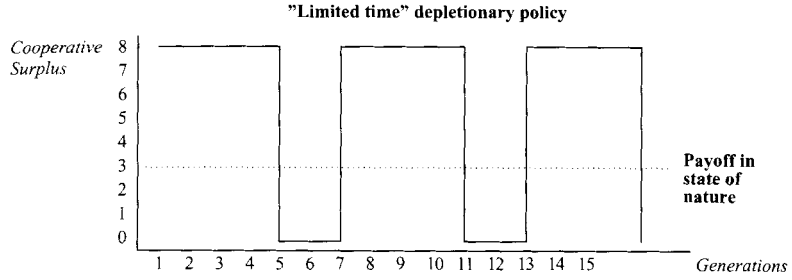

Figure 3

consequences of a policy stays in effect for one generation. Assume that the people in column 1 use a fertiliser that would increase the return in that game. In our example, the payoff for cooperation changes to 3x in column 1. In subsequent games, the output returns to normal (2x) until column 5 when the soil is useless and cannot produce any corn at all. From column 6 and onwards, however, the output returns to normal (2x) again. The only negative effect of the fertiliser is that the soil will be useless for the people in column 5. Then the following strategy is in equilibrium:

Strategy E: Always defect in a) the last game, b) in the last game that you can receive a cooperative surplus, and c) in games where there cannot be a cooperative surplus; cooperate in every other game until someone defects in a game that is not of type a, b or c for that person, then defect in every subsequent game.

The people in column 1 use a fertiliser that changes the payoff for cooperation to 3x for them and makes the soil useless for the people in column

	1 2000–20 (p1,p2,p3,p4)	**2** 20–40 (p2,p3,p4,p5)	**3** 40–60 (p3,p4,p5,p6)	**4** 60–80 (p4,p5,p6,p7)	**5** 80–100 (p5,p6,p7,p8)	**6** 100–20 (p6,p7,p8,p9)
Strategy E	(12,9,9,9)	(9,6,6,6)	(9,6,6,6)	(7,7,4,4)	(3,3,3,3)	(9,6,6,6)
p7 defects in C4	(12,9,9,9)	(9,6,6,6)	(9,6,6,6)	(7,7,2,**5**)	(3,3,**3**,3)	(3,**3**,3,3)
p6 defects in C4	(12,9,9,9)	(9,6,6,6)	(9,6,6,**6**)	(7,7,**5**,2)	(3,**3**,3,3)	(**3**,3,3,3)

Figure 4

tionary policies that stay in effect for two generations are not ruled out. Heath's model, where eight generations live at the same time, allows depletionary policies with negative effects that stay in effect for six generations. That this result is dependent on how we define a generation raises a problem which Heath does not discuss: How should we individuate generations in a non-arbitrary manner?

5 (that is, the payoff for cooperation is zero). If p7 defects in column 4, then he would receive a maximum payoff of 14 (5+3+3+3) as compared to 22 (4+3+6+9) if he follows strategy E. If p6 defects in column 4, then she would receive a maximum payoff of 17 (6+5+3+3) as compared to 22 (6+4+3+9) if she follows strategy E. In other words, a generation can adopt depletionary investment policies as long as the effects of such a policy are sufficiently bounded in time. Observe that this depletionary policy could be repeated an infinite number of times. For example, the people in column 6 could use the "depletionary" fertiliser again. Consequently, Heath's theory does not rule out depletionary policies.

Earlier generations can spend at the expense of future generations in another counterintuitive manner. It is enough that there will be a minimal cooperative surplus to secure cooperation with future generations.

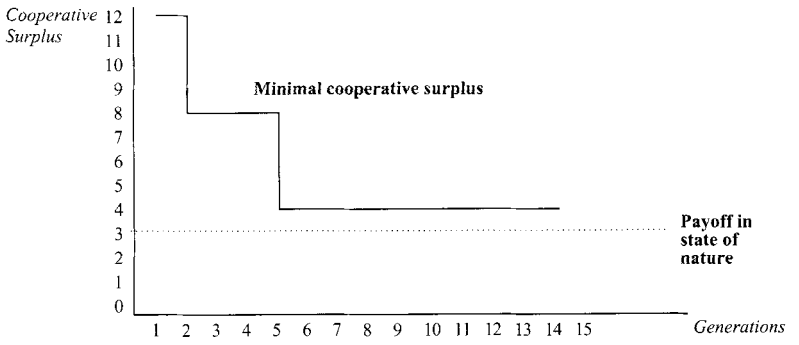

Figure 5

To use my example above, the people in column 1 could use a fertiliser that increases the return once but lowers the future payoff of cooperation in column 5 and onwards to 3+*m* (the payoff in the state of nature is assumed to be 3), where *m* can be an arbitrarily small positive real number (see figure 5).

Heath acknowledges that his backward induction argument does not eliminate inefficient equilibria.[9] He thinks that we can solve this problem by adopting a hypothetical intergenerational bargaining method as an equilibrium selecting device. This bargaining method would select

[9] Heath (1997), p. 370.

an equilibrium where, roughly, the cooperative surplus would be shared equally between earlier and later generations.[10] Since, according to Heath, any generation can terminate cooperation in all previous generations, "any investment policy must be one that each generation anticipates will be acceptable to all future generations. Naturally, the *minimal criterion* for the acceptability of an investment policy is that it not be depletionary. But there is nothing that stops us form imposing additional constraints on what policies are to count as acceptable. ... [S]ince game theory lacks an intentional equilibrium-selection mechanism, the instrumental conception of rationality does not impose any constraints on the reasonableness or credibility of the threat, nor does it dictate the criteria that determine what is to count as acceptable. This means that the theorist is free, quite literally, to make something up."[11] I find this reasoning quite odd. Firstly, what reasons do earlier generations have for *not* anticipating that future generations will cooperate in the "Limited time depletionary policy" and the "Minimal cooperative surplus" cases above? It would be irrational for future generations to defect in these cases, since they would get lower utility from defection than from cooperation. Secondly, if we appeal to considerations outside rational choice theory, the tier between morality and rationality is severed. We have then left the domain of mutual advantage contractarianism as it is usually conceived. Why not then directly appeal to the hypothetical bargaining procedure? Why do we then need any appeal to rationality when the hypothetical bargaining procedure is doing all the work in the end?

There is also an interesting and important difference between one shot games used by Gauthier and the iterated games used by Heath: In one shot games we can assume that the payoff in the state of nature is a constant. This is question begging for iterated games: We can surely affect the future payoff in the state of nature through, for example, environmental degradation (less fish in the river, less game in the forest).

Again, the people in column 1 could use a fertiliser that increases the return once but decreases the future payoff (column 5 and onwards) in the state of nature to 0 and the cooperative surplus to 1. The trigger strategy is in equilibrium in this game and although the policy adopted by the first generation is not depletionary according to Heath's definition, most of

[10] Ibid., p. 370–4.

[11] Ibid., p. 371 (emphasis in original).

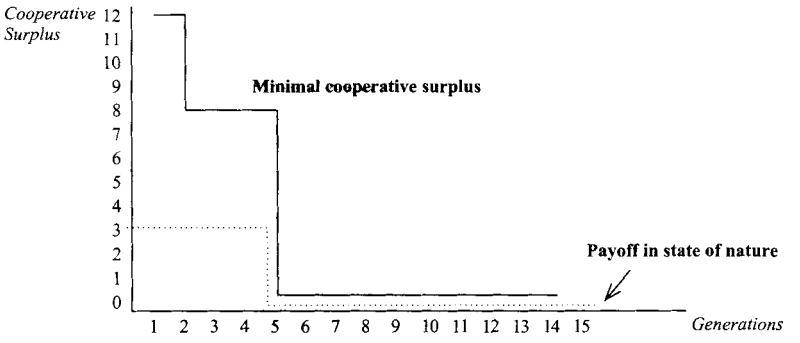

Figure 6

us, I suppose, would consider it as such. An interpretation of this game could be that future generations must cooperate in order to survive. It would indeed be odd if future generations would defect in such circumstances...

Conclusion

Heath has provided us with an interesting attempt to reconcile mutual advantage contractarianism with intergenerational justice. His model implies, however, that earlier generations, whenever they can benefit from it, ought to adopt depletionary policies at the expense of future generations. Intergenerational justice remains an embarrassment for contractarians and will, I surmise, continue to be so. The metaphor of a contract between mutually unconcerned individuals might be applicable to informed, competent adults. It is, on the other hand, somewhat bizarre to apply the contractarian metaphor to children and those yet unborn and other beings to whom we can only stand in an asymmetric relation of benevolence: One cannot get blood out of turnips.[12]

[12] I would like to thank Krister Bykvist, Erik Carlson, Bruce Chapman, Sven Danielsson, Adeze Igboemeka, Wlodzimierz Rabinowicz, Rysiek Sliwinski, Howard Sobel, Wayne Sumner and Jan Österberg for their comments and criticism on earlier versions of this paper. Earlier versions of this paper were presented at the Learned Societies Congress, Canadian Philosophical Association, Brock, June 1996 and Filosofidagarna, Lund, June 1997. I would like to thank the audience at these occasions for stimulating criticism. Financial support through a grant from the Swedish Institute during 1995–96 is gratefully acknowledged.

References

GAUTHIER, D., *Morals By Agreement*, Oxford: Clarendon Press, 1986.

HEATH, J., "Intergenerational Cooperation and Distributive Justice," *Canadian Journal of Philosophy*, Vol. 27, No. 3, September 1997, pp. 361–76.

RABINOWICZ, W., "Grappling with the Centipede," unpublished paper, 1996.

SLIWINSKI, R., "Baklänges induktion i spelteori," *Filosofisk Tidskrift*, nr. 4, Årg. 16, 1995.

SOBEL, H., "Backward induction arguments in finitely iterated prisoners' dilemmas: A paradox regained", *Philosophy of Science* 60, 1993, pp. 114–33.