

Making a Moral Corporation Artificial Morality Applied*

©Peter A. Danielson
Centre for Applied Ethics &
Department of Philosophy
University of British Columbia
Vancouver, Canada
Version 1.2

May 1, 1995

In my book, *Artificial Morality*, I defend a thesis and a method relevant to business ethics. The thesis is that moral constraint is rational; in some sense it does pay to be good. I sketch the argument for this thesis in section 1. Section 2 explains the method I employ: *moral functionalism*.¹ Section 3 applies my theory to the question whether a corporation can be moral using a management science fiction fable.

1 How it Pays to Be Moral

I argue that games like the Prisoner's Dilemma and Chicken model morally crucial unstable social situations. These are situations in which the agents — straightforward maximizers — recommended as rational by the received theory of rational choice cannot attain optimal outcomes.² I show how to build morally constrained agents — literally software robots — that do better than straightforward maximizers in these situations.

*Notes for a seminar on Business Ethics at U.B.C. Thanks to Michael McDonald for the invitation to present this material, to Hugo Chan for comments on an earlier version, and to Chris MacDonald for comments on this version, as well as the inspiration to move it to the Web. *Permission to reproduce this entire document for educational use hereby granted.* Comments welcome. Author's address: Centre for Applied Ethics, 227 - 6356 Agricultural Rd, Vancouver, B.C. Canada V6T 1Z2; email: pad@ethics.ubc.ca; FAX (604) 822-8627.

¹Section 2 is adapted from (Danielson 1992, §11.2). Hereafter I shall abbreviate this book title to *AM*.

²This thesis is a defense and elaboration and criticism of David Gauthier's theory of Morals by Agreement (Gauthier 1986) from which I take the categories of straightforward and constrained maximization.

1.1 The Extended Prisoner's Dilemma

Consider the simplest situation in which moral constraint make a practical difference, a one-play Prisoner's Dilemma (PD; see Figure 1). To simplify even

		II	
		C	D
I	C	2,2	3,0
	D	0,3	1,1

Figure 1: The Prisoner's Dilemma

further, let two agents, player I and player II, face an opportunity for *sequential* cooperation. (Figure 2 shows the decision tree for this extended form of the PD game.) If player I co-operates (*C*) and player II chooses *C* as well, both do better than if both defected (*D*). However, II can do even better if she responds to I's *C* with *D*; defection dominates cooperation.

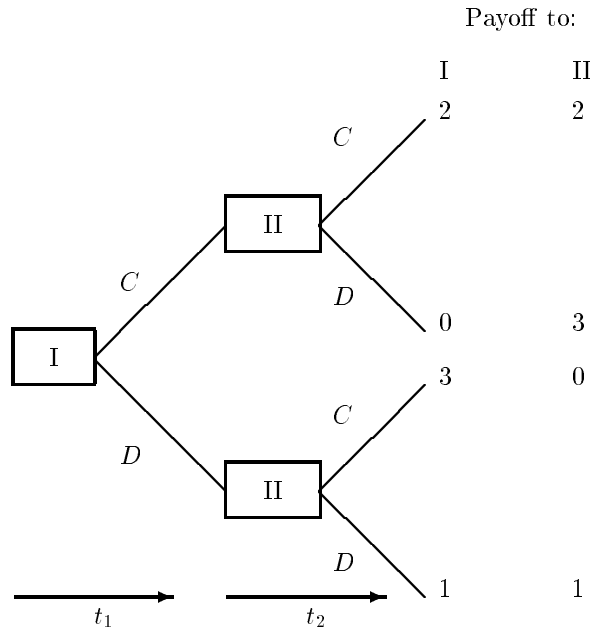


Figure 2: Extended Prisoner's Dilemma

Consider how a moral capacity could effect the situation. If player II is an amoral straightforward maximizer, she can be expected to choose the dominant strategy *D*, so player I should choose *D* as well. But if player II can be trusted to keep a promise to cooperate in response to (but only in response to) I's cooperation, I does best by choosing *C*. Therefore moral constraint — notice

that II must be able to choose, at time t_2 , what is *worse* (because she gets 2, not 3) — allows players to do better than unconstrained “rational” players. Of course the full story is more complicated.³ But this sketch indicates enough of the theory to support the central claim of what might be called instrumental ethics. In some crucial situations, discriminating morally constrained agents do better than unconstrained “rational” agents. It can pay to be moral.

2 Moral Functionalism

Artificial morality is functionalist. Given the goal of instrumental success, I ask what sort of entities best achieve this goal. My answer is that instrumentally rational agents must be capable of morality in the sense of a capacity to constrain their actions for the sake of benefits shared with others. In addition, their morality must be responsive; they must limit the class of agents with which they share co-operative benefits — although drawing the boundaries of their tolerance is a hard problem. For example, responsive players must be able to discriminate between cooperators and those that would exploit cooperative constraint. In *AM*, I argue that one can do this is by means of transparently public principles, which others can test and copy. In this way I proceed from the goal of instrumental success to responsive morality as general solution and Artificial Morality as a means to specify, implement, and test this solution.

If I am correct, the class of entities that can be moral agents is determined by their functional abilities. This basic functionalist doctrine is widely accepted for other capacities. For example, there are functional prerequisites for calculation. The fossil-filled stone I use as a paperweight would make a poor computer. It contains lots of silicon but not in a form that can change and retain states easily.⁴ Similarly, we may ask: What sorts of things make good — that is instrumentally successful — agents?

2.1 Half-Hearted Functionalism

Contractarians have always been somewhat friendlier to functionalism than other traditions in ethical theory. For example, Hobbes sees immediately that having created a new agent, the sovereign, his state of nature argument applies to these new agents (Hobbes 1968, Chap. XIII). Originally Rawls took a strikingly functional approach:

³For example, agents must be able to discern II’s disposition to co-operate, and interactions of various strategies safely to do this get complex. These are two reasons why it is helpful to build computer models of these agents and situations. *AM* works out the static modeling and (Danielson 1995) extends the account to allow dynamic, evolutionary models. Software and support for this research project is available on the Web; URL= <http://ethics.ubc.ca/people/faculty/pd.html>.

⁴A good introduction to the topic, which inspired this example, is (Haugeland 1987).

The term “person” is to be construed variously depending on the circumstances. On some occasions it will mean human individuals, but in others it may refer to nations, provinces, business firms, churches, teams, and so on. The principles of justice apply in all these instances, although there is a certain logical priority to the case of human individuals. As I shall use the term “person,” it will be ambiguous in the manner indicated (Rawls 1958, p. 166).

Gauthier also has applied his argument for constrained maximization to states as well as human individuals (Gauthier 1984). Nonetheless contractarians tend to half-hearted functionalism. The tradition barely tolerates non-human agents. Hobbes completes his second state of nature argument (applied to states) in a single sentence (instead of many pages) which is remarkable since it reaches a conclusion opposite from what the first leads one to expect (international anarchy instead of a super-national sovereign). Rawls has dropped the application of his theory to non-human agents and Gauthier grounds his rationality assumptions in appeals to rationality as essentially human (as I argue in *AM* Chapter 2). For the most part, contract theorists quantify over situations and principles, taking agents as fixed. They ask, In what situations do what principles best mediate co-operation for people? In contrast, *Artificial Morality* quantifies over types of agents as well, attempting to optimize over situations (games), principles, and agents.

I am a whole-hearted functionalist. I want to know what sorts of things are suited for useful moralized interaction. (What is required of agents and situations for moral agency to be an instrumental success? What moral principles suit these situations and capacities?) I was drawn to this functionalist method by its simplicity. It allows me to build (computer) models of whatever can be usefully moral without worrying about whether these models are satisfied by the usual subject of moral theorizing: human individuals.

2.2 Morality and Formal Organizations

It would be objectionable if *Artificial Morality* applied to nothing but little software robots. Is there any real thing that *Artificial Morality* is about? Yes; I think that it shed light on the controversial question of the moral standing of corporations. *Artificial Morality* may apply more readily to formal organizations because they have morally crucial capacities that people may lack. For example, firms and states may be constituted by public decision procedures, some of which commit them to various courses of action contrary to their interests, yet which are open to interest-based change.⁵ On the other hand, the scale of organizations and the fact that we can stand outside them (as well as function in them as

⁵I hasten to add the obvious qualifiers that not all firms and states are rational, open or adaptable. Cf. (Doyle 1983). Also, there are well-known problems in attempting to ground collective preferences to individual preferences, which I will not go into here.

components) allows us to see how something could make itself moral and more generally be moral using amoral parts. Since the idea of treating the state as a model of the individual is familiar from Plato's *Republic*, I should note a crucial difference. I will not claim that corporations are models of human individuals. I claim instead that they are models of adaptable rational agents, which, by satisfying my instrumental premise, are also capable of satisfying my moral conclusion. Whether humans can be so modeled remains an open question. I will argue that to the extent that formal organizations are like machines, Artificial Morality should apply to them.

2.3 Ladd's Objection

My analogy between organizations and machines is not original. For example, (Ladd 1970) invokes it to argue for a position quite contrary to mine:

... it will be evident from what I have said earlier that it is impossible to make a compact with an organization. (Can we make a compact with a machine?) A compact is a bilateral promise and hence a compact can be made only between beings that are capable of making promises. But a formal organization cannot make promises, for it cannot bind itself to a performance that might conflict with the pursuit of its goal. The principle of rationality, as applied to formal organizations, makes no provision for the principle that promises ought to be kept; indeed, if the keeping of promises, or of a particular promise, is inconsistent with the goals of the organization, that principle requires that they be broken (Ladd 1970, p. 504).

Ladd uses the analogy between machines and formal organizations to argue in the opposite direction from me. He assumes that it is obviously true that machines are incapable of moral constraint. Then he invokes the analogy between machines and formal organizations to conclude that organizations are incapable of entering moral relations. My argument critically addresses Ladd's assumption of machines' moral incapacity. Artificial Morality shows that when reduced to their functional minimum, moral relations are open to machines. The weakness in Ladd's argument is his failure to see that rationality is not limited to the direct connection between means and ends that I have characterized as straightforward maximization.⁶ Notice that I *agree* with Ladd that straightforward maximizers are not moral. But I have argued that for machines which face unstable social co-operation (e.g. play the Prisoner's Dilemma), well-designed principles of moral constraint are rational. In particular it is rational for machines to be able to keep promises (e.g. to co-operate with promise-keepers). Now I can invoke the machine/organization analogy to refute Ladd's conclusion. It is similarly

⁶In fairness to Ladd, I add that this assumption is widely shared by defenders of the received theory of rational choice.

rational for organizations to restrain themselves, hence corporations may enter into moral compacts.

Furthermore, contrary to Ladd's claim, there are reasons to think that among formal organizations, business corporations *in particular* would become constrained instead of straightforward maximizers. Firms are subject to strong evolutionary pressures, the so-called 'discipline of the market.' If I am correct that responsively moral agents are instrumentally more successful, this should be reflected in the outcome of the market selection process, to the extent that it links survival and substantial success.⁷ That is, market competition should select for moral (i.e. responsive co-operator) firms, not because their goals are moral (they need not be) but because some moral constraints are (indirectly) better means to market success. Ladd identifies his 'principle of rationality' with straightforward maximization on *a priori* grounds; I claim that this identification is refuted by this (conjectured) empirical market result.⁸ Therefore we have more reason to expect to find artificially moral firms in the market environment than to find amoral straightforward maximizing firms.⁹ I conclude that corporations provide an example of the applicability of my abstract instrumentalist moral theory.

3 How to Make a Moral Corporation

I turn to a second problem. What is the process by which an agent can change from straightforward to constrained maximization and how do principles motivate the resulting agent? So strong is the dominance of the received theory of rationality, especially in the case of corporate ideology, that it might be difficult for you to imagine a process whereby a corporation could be moralized. A little management science fiction fable might help. Let us imagine that the north american SM Corporation (SMC) finds it difficult to penetrate the Japanese market. Japanese suppliers are accustomed to long-term relations with purchasers; they say that the resulting trust lowers the cost of business. For example, they benefit by using fewer contracts and lawyers and they avoid price bargaining. SMC is a newcomer, a foreigner, and worst of all, the employer of the dreaded SM-MBAs, the coolest, most straightforward strain of maximizer yet. It is not surprising that SMC finds it impossible to assure its would-be Japanese suppliers.

⁷The whole discussion of corporate responsibility presupposes that market behaviour is not completely constrained by moral and legal rules, else it would be impossible for irresponsible firms to gain from force and fraud.

⁸I note several limits on this rough argument. Firms that are too small may not be able to adopt responsive moral methods. For example, a Mom & Pop candy store may be too closely identified with Mom and Pop to change its structure. At the other extreme, very large firms can influence their environment and so might be selected for the ability to corrupt political institutions rather than for market success.

⁹Of course, this conclusion depends on the market, legal, and political environment.

The management of SMC decides to change the nature of the corporation. They spin off a wholly self-controlled subsidiary, Hanko Corp. The guiding principle of Hanko Corp. is the Principle of the Hanko: the word of any Hanko manager, sealed by his personal corporate stamp (hanko in Japanese) binds Hanko Corp. Indeed, the managers make this rule the first operating procedure of the company, with strict priority over all other procedures.¹⁰

Now business proceeds smoothly. A manager, Peter-san, commits Hanko to buy a billion 256k-widgits. Given this assurance, Yokasuka Widgit is willing to commit itself to tooling up for this specialized product. Then one day one of the MBAs gets a bright idea. Now that Yokasuka Widgit is physically committed to large-scale production, why not pressure it to lower its price by threatening to move quickly to 512k-widgits? There is, after all, no legally enforceable contract between the two firms, only Peter-san's hanko. Doesn't the principle of rationality demand that Hanko defect from its informal promise in this case?¹¹

I claim that the MBA cannot change things and it is not rational that she change things. She cannot because the straightforwardly maximizing SMC no longer exists; in Hanko the management does not have the option to cancel the promised supply arrangement. This is the force of the Principle of the Hanko. Without this constraint, Hanko would not have been trusted. With it, it cannot defect. Note that I am not saying that it would be unfair; I am saying that it is not *corporately possible*. Of course the MBA could try to disturb this trust, but not everything that an employee of Hanko does is attributable to the corporation. The effect of the moralizing Principle of the Hanko is that Hanko Corp. is bound. Moreover, I claim that it is rational to be so bound. It is substantively rational, because this kind of bound agent does better in the set of unstable social interactions that constitute its ecological niche. If one replies that it remains rational to defect, because Hanko would do better thereby, one simply misreads the situation. The Hanko you are thinking about doesn't exist. The company that would be free to defect — SMC — is not and could not be in a *position* to exploit the trust of Yokasuka Widgit.

Conclusion

I conclude that I can imagine how a firm could become morally constrained even if it began as a straightforward maximizer.¹² This fictional example of

¹⁰Contrast my machine/corporation fable with the one in (Dan-Cohen 1986) where the corporation replaces men with machines. In my story the corporation *is* the (abstract) machine.

¹¹More subtle is the objection that straightforward maximization — not morality — demands promise-keeping in this case because of reputation. To meet this objection we need to further specify the example to block iteration. (Hanko is abandoning the island to move to mainland China?) The need for a real one-shot case drives one to the end of the world and problems like nuclear deterrence (Gauthier 1984). Note that a very big business deal can similarly swamp iteration's reputation effect.

¹²This argument fills out what I called the 'Constitutional Analogy' in *AM* Chapter 7, which used a political state rather than a firm. Firms are a better example than states because they

the morality of formal organizations illustrates my functionalist method and pragmatic approach to morality. If we do not prejudge what sorts of things might satisfy our accounts of rationality and morality, we may find interesting evidence that some forms of morality are rational policies for firms.

References

- Dan-Cohen, M. (1986), *Rights, Persons, and Organizations*, University of California Press, Berkeley.
- Danielson, P. (1992), *Artificial Morality*, Routledge, London.
- Danielson, P. (1995), Evolutionary models of cooperative mechanisms: Artificial morality and genetic programming, in P. Danielson, ed., 'Modeling Rationality, Morality and Evolution', number 7 in 'Vancouver Cognitive Science Series', Oxford University Press, New York.
- Doyle, M. W. (1983), 'Kant, liberal legacies, and foreign affairs, part I', *Philosophy & Public Affairs* **12**, 205–235.
- Gauthier, D. P. (1984), Deterrence, maximization, and rationality, in D. MacLean, ed., 'The Security Gamble: Deterrence Dilemmas in the Nuclear Age', Rowman and Allanheld, Totowa, N.J., pp. 101–22.
- Gauthier, D. P. (1986), *Morals by Agreement*, Oxford University Press, Oxford.
- Haugeland, J. (1987), *Artificial Intelligence: The Very Idea*, MIT Press, Cambridge, Mass.
- Hobbes, T. (1968), *Leviathan*, Penguin Books, Harmondsworth, Middlesex.
- Ladd, J. (1970), 'Morality and the ideal of rationality in formal organizations', *Monist* **54**, 488–516.
- Rawls, J. (1958), 'Justice as fairness', *The Philosophical Review* **LXVII**(2), 164–194.
- Werhane, P. H. (1980), 'Formal organizations, economic freedom and moral agency', *Journal of Value Enquiry* **14**, 43–50.

too have constitutions (Operating Procedures) and are directed to their amoral and external goal of profitability by strong pressures. In contrast, the goals of states are confused, morally infused, and internal. Hence I agree with (Werhane 1980) that clubs and nations are not like corporations but disagree with her conclusion that this disqualifies firms from following moral rules.