Ethics and efficiency in organizations

Kjell Hausken
Max-Planck-Institut für Gesellschaftsforschung, Cologne, Germany

Introduction

The article lays out a framework for analysing ethics in organizations. There are controversies in the intersection between economics and practical philosophy, and a need to fill the gap in the existing literature. The focus here is on the individual agent. Features in the surroundings are considered to the extent that these have significant bearings on our individual agent. We ask according to what criteria the individual agent should act, and what set of actions it is prudent for the individual agent to choose in the long run. The framework has five building blocks: self-interest, individual rationality, sequential rationality, incentive compatibility, and reputation. Self-interest may have negative social consequences. The article illustrates that an individual agent surprisingly often has an interest in constraining self-interested behaviour. Only those constraints that the individual agent agrees to by free will are accepted. Additional constraints involving altruism may benefit society, but are not treated here. Both within and outside the economics profession there is today a focus on the self-interest model, although some outspoken critics are hostile towards it. Uncritical training in the self-interest model may induce framing effects, as will be illustrated, blinding less cautious users to important ethical dimensions. As one's conceptions and ideas about how the world works may be affected by this model, one may ask what remedies there are to uncritical use of it. The article evaluates to what extent we can construct a remedy by starting out with the question, "What choice of strategy ensures the real interests of the individual agent in the long run?"

The fifth building block in the framework, the reputation concept, is broad, of a somewhat intangible nature, and can explain many different behaviours, perhaps too many. Our aim is to use it, and demonstrate that it can be successfully employed to a considerable extent in ethical analysis. We ask how far towards an "ethical world" one comes if each individual agent in choice of strategy carries out a cost/benefit analysis where also reputation is a relevant input. Is it rational for an individual agent to constrain self-interested behaviour further? Can a society expect further constraint? Observe the emphasis of the time aspect in the framework, contrary to most alternative ethical theories ignoring the time dimension. The three concepts – sequential rationality, incentive compatibility and reputation – all involve considerations over time. The presence of the time aspect enables us to assign a more solid and fundamental ethical "touch" or "flavour" to our framework than what is possible in a purely static analysis.

International Journal of Social Economics, Vol. 23 No. 9, 1996, pp. 15-40. © MCB University Press, 0306-8293

Approaches like this article elicit questions. What is the objective of ethical analysis? Is there an attempt to legitimate economic theory as ethical? Is there an attempt to teach or preach to economists or ethicists how to become more ethical? Do we seek to raise ethical awareness against economists' or ethicists' own interests? The response here is to produce no unjustified normative guidelines, except those to which "economic woman" finds an interest in assenting by free will. More specifically, we pose economic woman, and teach her to expand on narrow shortsighted self-interest along two dimensions. First, we teach her to focus on her real and actual interests in a broad sense. Second, we teach her really to think long term. For example, if acting virtuously contributes to a character or personality which subsequently and indirectly influences economic woman's reputation beneficially in the long run, then this action is to be recommended if the long-term benefit of the beneficial reputation outweighs the short-term benefit of a more deceitful or vicious action. In other words, our "ethical economic woman" is a trade-off talking rational economic person, constantly carrying out a cost/benefit analysis where reputation is a relevant input, constantly focusing on her real interests in a broad sense, and constantly thinking long term. Thus ethical economic woman is a normative ideal. She is an external standard according to which we assess ethical behaviour. We may label this standard a necessary minimum requirement for ethical behaviour in general. Actions falling short of this standard we will label unethical. Actions going beyond the standard we will not take into consideration. Such actions may be ethical according to other ethical theories, prescripts, recommendations, etc. They may be beneficial in a collective sense, for a community as a whole, and perhaps even for a person's "inner personal life", etc. Nevertheless, such forms for ethical behaviour are disregarded and not recommended because they are not in the interests of economic woman, since they do not contribute even indirectly to a beneficial reputation. That is, our ethical economic woman has no soul except in so far as having a soul gives reputational effects.

There are at least three potential benefits of the approach. The first is to illustrate to the individual agent where her real interests actually lie. Visualizing various aspects of self-interest to the individual agent may give her a richer and more solid basis from which to choose strategy. The second potential benefit follows from the first in that it may increase our understanding of how to structure the surroundings around the individual agent. How to do this structuring lies beyond the scope of this paper, and will only be treated briefly. The third potential benefit follows from the first two, but is more general. It has to do with my belief that making mechanisms, procedures, and underlying assumptions explicitly pronounced, increasing insight and common knowledge, illustrating key aspects in the nature of human interaction, etc., is beneficial from an ethical point of view.

In the next three sections the first three building blocks are introduced. The following section describes incomplete information, which is then used in the introduction of the last two building blocks. The article then considers the relationship between principal-agent theory and ethics, and follows this by illustrating how individual behaviour is embedded in a surrounding structure. Finally the article analyses the real interests of the individual agent in the long run. Observe that our approach is firmly rooted within the economic paradigm. This is because I believe that the main ethical challenge today is not *homo sociologicus*, but rather *homo economicus*, and because one effective way of analysing economics involves using economic tools. Thus the approach is compatible with Taylor's (1987, p. 30) suggestion that "the 'solution' of collective action problems by norms presupposes the prior or concurrent solution of another collective action problem", which is slightly at variance with Elster (1989, p. 15), who has "come to believe that social norms provide an important kind of motivation for action that is irreducible to rationality or indeed to any other form of optimizing mechanism". This does not negate the plausible view that behaviour based on calculation over time may turn into unreflective automatic habits, or perhaps even that our very actions originally stem from other principles than optimization. Nevertheless, analysing behaviour in terms of calculation is effective and serves our purpose in this article, in addition to conforming with basic notions, values, and principles regarding how modern woman functions in contemporary society.

Self-interest and other-regarding behaviour

Hume (1751, app. II) criticizes Hobbes and Locke for "explaining every affection to be self-love, twisted and moulded, by a particular turn of the imagination, into a variety of appearances", and "esteems instead the man whose self-love, by whatever means, is so directed as to give him a concern for others, and render him serviceable to society". This dispute is partly about words. However, to the extent that there is real difference between the two views, we should admit that both self-interest and sympathy may be motives or actuating principles in human nature. They may both be present in a higher or lower degree, they may work in tandem, and they may both have explanatory power in different contexts. Economic theory has a natural connection to Hobbes' self-interest model. Other academic disciplines analysing driving forces in human nature may embrace other models. We shall later see to what extent potential negative sides of the self-interest view can be countered by the reputation aspect. If sympathy is actuating in human nature, and to the extent that transmitting sympathetic feelings gives reputation effects, the sympathy aspect is present in our framework. A key question, however, is in what contexts the self-interest model or the sympathy model is to be used.

In negotiations where economic aspects are crucial and where each agent acts as a representative for specific interests, a self-interest model is likely to have greater explanatory power than a model with focus on other-regarding behaviour. Behavioural interactions and negotiations between employers and employees regarding, for example, how to increase a pleasant atmosphere in the organization can of course be explained by other models. One might venture to assume that to the extent that employers invest considerable resources to increase a pleasant atmosphere, this is an example of other-regarding behaviour. However, this behaviour can also be explained by the self-interest model.

Making a pleasant atmosphere can be considered as part of compensation to employees, and it may contribute to the establishment of a corporate culture giving various kinds of positive synergy-effects; increased effort level, co-operation, etc. The self-interest model is crucial in economic theory. It has explanatory power as a descriptive theory. To the extent that the theory has normative implications, our problem is to learn to use it in the correct way.

Individual rationality

Game theory assumes that each individual agent's decision-making behaviour is rational if it is consistent with maximization of subjective expected utility. One typically assumes Cardinal von Neumann-Morgenstern utility, and allows for interpersonal comparisons of utility. Individual rationality implies that every agent wishes to play the game. In principal agent analysis a better name for the individual rationality constraint is participation constraint; the principal makes it individually rational for the agent to participate by compensating the agent enough so that there do not exist outside opportunities which the agent prefers.

We trivially observe that satisfying the individual rationality constraint implies rejection of utilitarianism. Rawls' (1971) society "is a cooperative venture for mutual advantage". His lexical difference principle ensures that no one's interests are sacrificed in order to improve the lot of someone better off. However, the principle does not necessarily command everyone's support. The subject acting in Rawls' moral world is the Kantian "real self", revealed by removing all knowledge of contingent features of individual identity. Thus Rawls collectivizes individual assets (compare Gauthier, 1986, p. 254), thereby legitimating free-riding and parasitism. A free-rider obtains a benefit without paying all or part of its cost. A parasite in obtaining a benefit displaces all or part of the cost on to some other agent. Hence Rawls does not satisfy the individual rationality constraint. Gauthier (1986) denies the existence of a Kantian "real self" and claims that individual identity in all respects is contingent on social conditions

Gauthier's Archimedean point is similar to Rawls' veil of ignorance regarding focus on impartiality. Gauthier gives an individual agent a right to her factor endowments but assumes impartially constrained behaviour in ignorance of how one actually is factor endowed. He argues extensively that such behaviour is individually rational. "Proofs" here are of course not easy to furnish. Arguments very easily take the following somewhat loose form (Gauthier, 1986, p. 265): ignorance of one's own identity "is appropriate to the selection of those principles embodied in the institutions and practices that constitute our social womb … In identifying with the ideal actor, we justify to ourselves the social processes by which we have come to be".

Sequential rationality

Sequential rationality is the most crucial component in a sequential equilibrium (Kreps and Wilson, 1982a), today's most widely used equilibrium concept for extensive games and repeated games with incomplete information. Most ethical theories are of rather static nature. Gauthier (1986) suggests a two-step procedure by demand in the first stage, followed by concession to the "ethical" solution in the second stage. Observe that these two steps can be carried out simultaneously in time. The time aspect is of crucial importance. For example, giving a promise today with the intention of breaking it tomorrow necessarily involves various kinds of considerations over time. The main idea regarding sequential rationality is that every decision must be part of an optimal strategy for the remainder of the game. That is, the strategy of each player starting from each information set must be optimal, starting from there according to some assessment over the nodes in the information set and the strategies of everyone else. A player must not, after having established her beliefs, at any time at any node in the extensive form game, wish to change her strategy if this strategy is to be sequentially rational. For mechanisms involving several agents, sequential rationality requires that it must never be common knowledge that the mechanism induced over time is dominated by an alternative mechanism.

Incomplete information and type theory

Before we introduce the next building block, incomplete information and the Harsanyi (1967-68) doctrine (his type theory) must be introduced. Games with complete information presume that all players know all strategies available to all players and all potential outcomes for all players. If some of this is not fulfilled, the game has incomplete information. The break-through for analysing such games came with Harsanyi (1967-68). He assumes that each player has a subjective probability distribution over the alternative possibilities for the incomplete information. Our main problem is how to formalize this incomplete information. The next problem is to specify how the players construct their subjective probability distributions, i.e. how they form/update their beliefs and conjectures. Harsanyi assumes in his three articles (1967-68) that the probability distributions to the various players, i.e. their beliefs, are mutually consistent, i.e. that they can be derived from Bayes' rule from a common joint prior distribution over the unknown parameters for the various players.

The mutual consistency assumption is necessary to use the revelation principle, by ensuring that there is a probabilistic way in which nature can move. Normally one assumes that this prior common distribution is common knowledge among the involved players, but not necessarily public knowledge in the sense that outsiders are also informed. This is, then, the *ex ante* situation. With the assumption of mutual consistency the original game can be substituted by a game where nature makes the first move in a lottery according to the prior commonly known distribution. The result of this lottery, about which each player is only partly informed, determines which new game is to be played, i.e. with what parameters the original game is to be played. The new game is called the Bayes equivalent to the original game, and is technically a game with complete information. The situation before the new game is played is called the interim situation, in that each player has some private information, in addition to the common knowledge about prior distribution. The *ex post* situation ensues when all private information is known. Consider the type concept.

Harsanyi (1967-68) assumes that all players know all the strategy spaces to all the players. This will often be a realistic assumption. However, the justification is that the eventual lack of information one player might have about another player's strategy space may be represented in the model as lack of information about the latter's pay-off, e.g. by assigning low pay-off to unlikely strategies. In other words, no generality is lost by the assumption of common knowledge about strategy spaces. Harsanyi's main aim is to avoid the complicated formalization of the following problematic infinite chain of reasoning: "If I think that you think that I think that you think that…".

Briefly, in a two-person game player *i*'s *k*th-order expectation ($k > 1$) will be a subjective $P_i^k(P_j^{k-1})$ over all alternative ($k - 1$)th-order subjective probability distributions P_j^{k-1} that player *j* may possibly entertain. With several players the formalization is more complex. Harsanyi assigns a so-called attribute vector to each player *i*, describing players *i*'s private information. This attribute vector represents certain physical, social, and psychological attributes of player *i*, in that it summarizes some crucial parameters of player *i*'s own pay-off function as well as the main parameters of his beliefs about his social and physical environment. In other words, each player *i* is now allowed to belong to any of a number of possible "types", corresponding to the alternative values her attribute vector could take.

Consider an example. The seller of a house may be uncertain regarding the buyer's reservation price, i.e. the highest price the buyer might be willing to pay. The seller, then, can consider each possible reservation price of the buyer as corresponding to a specific type of this buyer. Alternatively, the incomplete information may be formalized as uncertainty about the buyer's time preference, hence different discount factors may correspond to different types for the buyer.

There is, then, a continuum of types. The subjective probability distribution one player assigns to another player's private information describes how likely this other player is to be of one or the other type. By this formalization one avoids forming probability distributions over probability distributions. We have instead probability distributions over types, and each player updates her subjective probability distributions or beliefs about the other players' types over time, depending on how the game proceeds.

Rather than relying on the Harsanyi doctrine, Gauthier (1986) distinguishes between constrained maximizers (CMs) and straightforward maximizers (SMs). In short, and simplified, a CM chooses to co-operate in a prisoner's dilemma situation if she expects the opponent to co-operate, but not otherwise (Gauthier, 1986, p. 167). A SM, on the other hand, simply maximizes her utility given the strategies of those with whom she interacts. The closest Gauthier comes to our type-concept is by his (1986, p. 167) observation that actors may be, as he calls it, transparent, translucent or opaque. A transparent actor's type, CM or SM, is fully known by other actors. Translucency means that one's disposition to co-operate or not may be ascertained by others, not with certainty, but as more than mere guesswork. Translucency is a more realistic description than transparency. We trivially observe that translucent CMs must expect to do less well in interaction than would transparent CMs, whereas translucent SMs must expect to do better than would transparent SMs. Hence, if one is transparent, it is rational to choose to be CM. If one is only translucent, choice of strategy is more complicated. A high percentage of CMs in the population implies higher benefits from choosing constrained maximization. A lower percentage of CMs increases the possibility of being exploited by SMs. Gauthier's theory is mainly static. Although taking a Hobbesian (Locke-inspired) game-theoretic approach, he incorporates little of modern game theory after Zeuthen-Nash-Harsanyi, e.g. pertaining to incomplete information and extensive form games.

Incentive compatibility

A main critique against most of today's ethical theories is that they are mainly static, and do not preserve the incentive compatibility constraint. Especially, an agent has to be preserved with incentives inducing truth telling, i.e. implicitly in carrying out an action an agent will reveal something about her own type, beliefs, preferences, etc. Consequently, in choice of strategy (i.e. a complete plan of action; what to do in every contingency) the agent will have to take into account not only what she might know about other actors' types and the structure of the game, but also her own type. The idea is that revealing one's own type may make one vulnerable and a potential victim of exploitation. Certainly, this is not always the case. Sometimes an individual agent has an interest in disguising her type. Other times the circumstances may be more fortunate, and the agent may actually have an incentive in ensuring that her type is revealed properly, to ensure that unfortunate consequences of false identification do not result. Equilibria in games of this kind are called screening or separating equilibria, where a dominant motive for the actors is self-selection. A main problem from an ethical point of view is not when actors reveal information about themselves, but rather when they disguise their own type. Having a motive to disguise one's own type means that one has a motive to imitate another and more "beneficial" type, i.e. act as if one were of another type and thereby potentially signal more "beneficial" information about oneself to the other players in the game. A crucial question is thus whether an agent has incentives to choose actions revealing knowledge about her own type. The incentive compatibility constraint is said to be preserved if the agent is induced to act in a way revealing true information about her own type.

Consider an adverse selection problem with hidden information in principal-agent analysis. The agent may be of two types, t_1 or t_2. The risk-neutral principal assesses a probability p_i, i = 1, 2, to the event that the risk-averse agent is of type t_i, i = 1, 2. Let x_i denote the amount produced by agent i, and s_i the amount paid to agent i, i = 1, 2. Agent i's pay-off is $u(s_i)c(x_i, t_i)$. Assume that c is convex, $c(0, t_i) = 0$, and $c(x, t_2) < c(x, t_1)$ for all $x > 0$, i = 1, 2, so that type t_1 has a higher cost of production than t_2. The principal seeks to maximize

$$\max \{ p_1 (x_1 - s_1) + p_2 (x_2 - s_2)\} \tag{1}$$

subject to

$$\text{IR1: } u(s_1) - c(x_1, t_1) \geq u = 0 \text{ (w.l.o.g)} \tag{2}$$

$$\text{IR2: } u(s_2) - c(x_2, t_2) \geq u = 0 \text{ (w.l.o.g)} \tag{3}$$

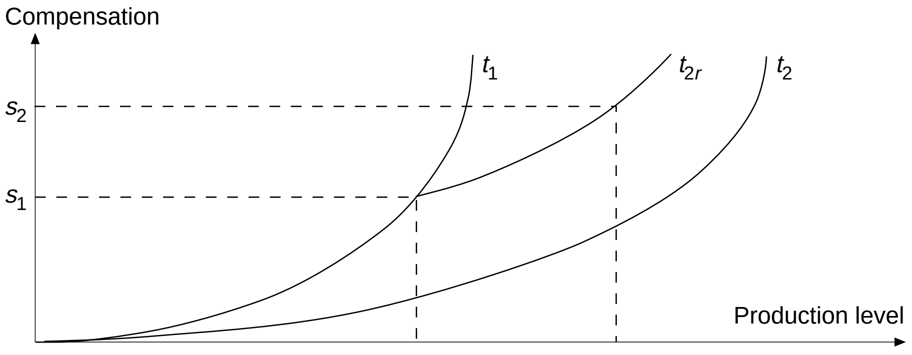
$$\text{IC1: } u(s_1) - c(x_1, t_1) \geq u(s_2) - c(x_2, t_2) \tag{4}$$

$$\text{IC2: } u(s_2) - c(x_2, t_2) \geq u(s_1) - c(x_1, t_2). \tag{5}$$

At least one of the individual rationality constraints (participation constraint is a better word) IR1 and IR2 must be binding, since the principal otherwise can raise profit by lowering both s_1 and s_2 without changing the increment $u(s_2) - u(s_1)$. The assumption $c(x_1, t_2) < c(x_1, t_1)$ together with equations (5) and (2) give

$$u(s_2) - c(x_2, t_2) \geq u(s_1) - c(x_1, t_2) > u(s_1) - c(x_1, t_1) \geq u \tag{6}$$

which implies that if IR1 is satisfied as an inequality, then IR2 is satisfied as a strict inequality. Hence IR1 is binding, and IR2 not. In other words, the principal is advised to construct a compensation scheme where the individual rationality constraint is binding only for that agent type (t_1) having the highest cost of production. To provide intuition for how to preserve incentives, consider Figure 1. The indifference curve for type t_1 is steeper than for type t_2 due to type t_1 having a higher cost of production and hence needing higher compensation for a given production level. The principal can induce type t_1 to produce x_1 by compensating s_1. Figure 1 shows that the production level/compensation pair (x_1, s_1) is more beneficial for t_2 than for t_1.



Note: Curve t_{2r} corresponds to curve t_2 removed along the vertical axis so as to go through point (x_1, S_1)

Figure 1. Indifference curves for compensation as a function of production level for agent types t_1 and t_2

Hence type t_2 may have an incentive to disguise or masquerade as being of type t_1. Hence our crucial question becomes: how can the principal induce type t_2 to produce x_2? Focusing solely on the indifference curve for type t_2 is insufficient, since the principal has imperfect knowledge whether the agent is of type t_1 or t_2, and t_2 may have an incentive to disguise as being of type t_1. Taking this latter fact into account we remove the indifference curve for type t_2 along the vertical axis in Figure 1 so as to go through point (x_1, s_1). Denote this removed indifference curve as t_{2r}. Hence the principal must compensate s_2 in order to induce type t_2 to produce x_2. In other words, the agent of type t_2 is compensated so as to be indifferent between (x_1, s_1) or (x_2, s_2), thereby providing type t_2 with incentives to choose a production level revealing true type (t_2) rather than disguising as being of type t_1. From Figure 1 observe that type t_1 will not have any incentive to disguise as being of type t_2. It is that type (t_2) having, for the principal, a more beneficial indifference curve which may have an incentive to disguise as being of a type (t_1) having a, for the principal, less beneficial indifference curve. This indicates that IC2 is binding, and IC1 not. Hence remove IC1, i.e. equation (4), from the principal's maximization programme. We will afterwards show that IC1 is satisfied. IR2 not binding implies that equation (3) can be removed from the programme. IR1 binding implies that equation (2) can be written as $u(s_1) = c(x_1,t_1)$, which inserted into equation (5) gives

$$u(s_2) = \mathrm{dc}(x_1) + c(x_2, t_2) \tag{7}$$

where

$$\mathrm{dc}(z) = c(z, t_1) - c(z, t_2) > 0. \tag{8}$$

Note that $\mathrm{dc}(z)$ is increasing in z. Let f denote the inverse function of u, i.e. $f = u^{-1}$. The principal's maximization programme can thus be written as

$$\max \{p_1[x_1 - f(c(x_1, t_1))] + p_2 [x_2 - f(\mathrm{dc}(x_1) + c(x_2, t_2))]\}, \tag{9}$$

where s_1 and s_2 are substituted away, implying no presence of constraints. The first-best solution to this maximization problem involves choosing x_1 and x_2 so that

$$x_1^* = \mathrm{argmax}[x_1 - f(c(x_1, t_1))], \tag{10}$$

$$x_2^* = \mathrm{argmax}[x_2 - f(\mathrm{dc}(x_1) + c(x_2, t_2))]. \tag{11}$$

Since f is an increasing function, the optimal (second-best) solution for the principal is to choose $x_2 = x_2^*$ and $x_1 < x_1^*$. Finally show that IC1, equation (4), is satisfied. Rewriting equation (4) gives

$$c(x_2, t_1) - c(x_1, \mathrm{t}_1) \geq u(s_2) - u(s_1), \tag{12}$$

into which insertion of equation (5) as an equality, corresponding to IC2 binding, gives

$$c(x_2, t_1) - c(x_1, t_1) \geq c(x_2, t_2) - c(x_1, t_2), \tag{13}$$

i.e. $\mathrm{dc}(x_2) \geq \mathrm{dc}(x_1)$, which is satisfied since $x_2 = x_2^* > x_1^* > x_1$ and since $\mathrm{dc}(z)$ is increasing in z. Hence we have shown that the principal has an interest in constructing a compensation scheme ensuring participation by the agent of type t_1, while inducing the agent of type t_2, i.e. the agent with high ability and low cost of effort, to choose a production level which truthfully reports this agent's type. In general, this solution is such that resources from the agents having high ability (t_1) are exploited in an optimal way, whereas agents with lower ability (t_1) are induced to choose a lower effort level.

We have criticized ethical theory for not embodying the crucial incentive compatibility constraint. Now evaluate possible implications of incorporating this concept. We have observed that Rawls (1971) and Gauthier (1986) assume some degree of impartiality in that one is assumed to act in ignorance of one's own identity, thereby implicitly choosing a social structure consisting of persons individuated in all possible ways. Acting like this is consistent with the most central principle in bureaucratic theory, the principle of universality. In our culture, laws, and customs, aspects like equality before the law, avoidance of indifferent treatment with insufficient justification, etc. have importance. At least, such universal values function as legitimation devices, according to which we justify more dubious behaviour, thus potentially indirectly having disciplining influence. We may ask whether agents have incentives to act in this "universal" way. Schmidtz (1989, ch. 8) states that "the norms people should choose for themselves differ from what would be right if they were choosing for everyone". There is an inevitable trade-off between economic and ethical theory in this regard. Consider four steps in ethical analysis, and then evaluate possible implications of our fourth step: utilitarianism; Rawls' (1971) theory; Gauthier's (1986) theory; and this article's framework. Rawls (1971) criticizes utilitarianism for not taking seriously the distinction between persons. Gauthier (1986) criticizes utilitarianism and Rawls' theory for not taking seriously the individuality of persons, and especially Rawls' theory for collectivizing assets. Gauthier (1986) seems to preserve the individual rationality constraint. He takes one further step towards economic theory by giving each agent the right to her factor endowments. He is fully aware that this implies some de-emphasis of the central value equality. One hypothesis is that incorporation of incentive preservation in ethical theory may further increase inequality, although such a hypothesis is hard to evaluate. Consider possible implications of our fourth step. Traditional ethical theory focuses on equality because we hesitate in incorporating notions of differences in order of rank, and other differences among persons. However, two kinds of differences that are difficult to avoid are factor endowments and ability.

Let factor endowments refer to goods, wealth, how well one is equipped and endowed aside from ability, etc. Factor endowments in negotiations may refer to the status quo point, i.e. one's utility level while negotiations proceed or the utility level reached if negotiations break down and one has to seek the best outside opportunity. Giving actors right to their factor endowments involves legitimating those assets which are brought to the bargaining table. By ability we seek to emphasize what we illustrated in Figure 1.

Actors may have different costs of production, some agents may have better ability or be more talented so as to produce a given amount x by investing lower effort, and some actors may be so intrinsically satisfied with their work that they need almost no compensation, etc. Given differences in factor endowments and ability among actors, the challenge is how to build out an ethical theory embodying such differences, while accounting for hidden action and hidden information involved. Returning to our equality hypothesis above, we saw in our principal-agent example that the agent with low ability (t_1) had to be paid more than the agent with greater ability (t_2) to ensure participation. However, this latter agent had to be paid more to preserve incentives, and the solution suggested efficient use of this latter agent, whereas the former was utilized to a lower extent. Hence consider a second hypothesis stating that preserving the incentive compatibility constraint leads to increased differentiation. Differentiation and specialization are crucial in today's society and may be beneficial from an efficiency point of view. Evaluating negative sides of differentiation is complicated. Agents utilized to a low degree may be thereby induced to specialize and qualify for more suitable work.

We have now introduced the first four basic building blocks in our framework. A bargaining scheme satisfying individual rationality, incentive compatibility, and sequential rationality is called a perfect bargaining mechanism. In a sense this is the hard rock bottom from which we started out in our approach. Our next step is to incorporate an ethical element into our framework. We do this by focusing on the individual agent, and not bringing in other considerations than those corresponding to this actor's real interests in the long run.

Reputation

Already Hume, 250 years ago, was aware of the importance of the reputation concept. Short and direct, he simply says (1740, p. 501, section 3.2.2): "There is nothing, which touches us more nearly than our reputation". There are various concepts or words with ethical "odour"; reputation, commitment, credibility, etc. We have decided to focus on reputation, but it is clear that this concept is interdependent with other concepts, e.g. with commitment. To enable a role for reputations one must assume (Wilson, 1985, p. 29; see also Kreps and Wilson, 1982b) several players in the game, that at least one player has private information that persists over time, that this player is likely to take several actions in sequence, and that the player is unable to commit in advance to the sequence of actions she will take. In a sense, commitment involves choosing an action, and then "burning one's bridges" (Schelling, 1960), thereby ensuring some degree of irreversibility. Commitment implies that one has placed restrictions on oneself. To be of any use, the commitment must be known by other players. It is clear that being able to commit oneself publicly in advance to this or that strategy may be beneficial for an individual agent. The crucial problem is how the individual agent is to make the commitment trustworthy, i.e. ensure that she is believed. To the extent that an agent is able to commit in advance to a certain strategy, her behaviour will be perfectly predictable. If one's strategy is perfectly predictable, one's reputation is irrelevant, since neither one's own nor others' strategies depend on this reputation. Thus, ensuring a role for reputation is done by introducing some degree of uncertainty regarding how the agent will behave in the future. That is, at a later stage in the game the agent will consider the matter anew and choose an action that is part of an optimal strategy for the remainder of the game. In other words, at this later stage our agent will reinitialize the subgame that remains by taking her reputation and the reputations of others as the initializing probability assessments that complete the specification of the subgame. This illustrates that there will be trade-offs between short-term consequences and long-term reputational effects of actions taken in earlier stages of the game. In situations where imitation through disguise of one's own type is the dominant motive, there will be a trade-off, in a multi-period game, between choosing an action which in some sense is beneficial, and which at the same time does not reveal one's own type in an unfavourable way. Illustrate this phenomenon by the well-known chain-store game. There are potential entrants to the market. Assume potential entry in sequence. In each period, if the entrant stays out, the incumbent's pay-off is 8. If the entrant enters, her (the entrant's) pay-off is 2 or –2 depending on whether the incumbent fights or acquiesces. (In acquiesce) gives pay-off 4 to the incumbent, whereas (in-fight), implying, for instance, that the incumbent starts a price war by pressing down prices, gives the lower pay-off 2 to the incumbent, in this period. In a one-shot game we see that the unique sub-game perfect equilibrium is (in, acquiesce) giving payoff (2, 4). In a finite-horizon game this equilibrium is sustained in the last period.

Hence, the incumbent will not benefit from fighting in the next to last period because the only possible reason for doing so is to avoid entry in the last period. By backward induction, this argument applies all the way up to the first period. With an infinite sequence of entrants there is no "last period", so consider this case. Assume that the incumbent discounts future pay-offs according to the discount factor δ. By fighting in the first period the entrant may acquire a reputation for fighting, hence avoiding entry in subsequent periods, thus receiving pay-off 8 in future periods. However, fighting in the first period gives pay-off 2, whereas acquiescing gives 4. This gives an infinite geometric series, which implies that it is rational for the incumbent to fight if $2 + 8/(1 - \delta) \geq 4/(1 - \delta)$, i.e. if $\delta \geq 1/3$. Therefore, the incumbent's time preference, i.e. how future is discounted, is crucial and it will be rational for the incumbent to nurture a beneficial reputation (for fighting) if future is sufficiently important. We see that there is an inherent instability in this game. What if the incumbent is weak, and cannot afford a price war, but nevertheless attempts to signal to potential entrants that every entry will be fought? Will the entrant accept the "bluff" from the potentially weak incumbent? What this game may potentially also illustrate is the crucial feature of self-selection, briefly mentioned in the last section. If the strong incumbent is able to take steps that the weak incumbent under no circumstances is able to take, i.e. imitation is impossible, the strong incumbent will be in the potentially favourable situation of being able to select itself out, thereby ensuring correct identification by outsiders.

Consider another example, from bargaining theory. The seller of a house (reservation price being type) will, assuming repeated offers/counter-offers over time, in each stage of the game choose between demanding high price (potentially signalling high reservation price (type), potentially obtaining high surplus, but also potentially delaying or failing to reach agreement) and demanding low price (signalling low reservation price, increasing probability of agreement, though with potentially smaller surplus). Modern bargaining theory (e.g. Cramton, 1988) assumes that each agent assigns a subjective probability distribution to other actors' types. Relating to our reputation concept, the seller's reputation before her offer is identical to the buyer's probability assessment of the seller's valuation, hence contingent on past history (offers, rejections, etc.). Both bargainers, then, take their own and their opponent's reputation into consideration, and choose an action as part of an optimal strategy for the expected remainder of the game, depending on what inferences the opponent is likely to make about one's own type, and depending on how one is likely to act oneself at a future time. Thus, the player's strategy is the solution to a dynamic programming problem in which both bargainers' reputations are among the state variables that link successive stages of the game. Hence, negotiations can be considered as the evolution of the actors' reputations. Considering bargaining theory from an ethical point of view a crucial question is whether the structure of the game and the situation broadly considered is such that the actors are induced to report their types in ways that are "ethically beneficial" in some sense, e.g. collectively optimal. To the extent that truthful reporting of one's own type is ethically beneficial, incentive compatibility is an ethically useful concept.

One may argue against our use of the reputation concept that sometimes the best reputation to have is not a reputation for focusing on ethics, but a reputation for being a bully. A response to this is that, to the extent that being a bully implies not being ethical, the remedies for avoiding potential harmful effects of acting as a bully are to be constructed in the surroundings around our actor. Such analysis lies beyond the scope of this article.

Principal-agent analysis and ethics

The self-interest model is essential in many economic models, e.g. bargaining theory, auctions, principal-agent analysis. One appropriate candidate posing a challenge is principal-agent analysis, which embodies the four basic concepts in our framework and especially emphasizes the incentive compatibility aspect. Taking the individual agent as our starting point, we seek to analyse what kind of reputation our agent finds it beneficial to nurture, in relation to other actors, and surrounding structure in which the agent is operating. We seek to evaluate to what extent the reputation concept can be used, and how far one can go with it. In a sense we attempt to take a first step, by focusing on the individual agent, in establishing an ethical framework or theory of the firm. The second step, not taken in this paper, consists of analysing the surroundings around the agent, i.e. the firm in general.

Principal-agent analysis usually makes four assumptions, consonant with the four basic concepts in our framework: goal incongruity (although perhaps finding the work intrinsically satisfying, it is assumed that the agent to some extent must be induced to work, by monetary compensation, by the presence of a pleasant work-atmosphere, etc.); uncertainty (observed outcome may depend on a variety of factors, including an actors effort level); asymmetric information (knowledge, normally by the principal, only about a probability distribution of the agent's type); and risk aversion (the agent prefers a given compensation with certainty, rather than an infinitesimal higher compensation with uncertainty, i.e. depending on factors beyond the agent's control). The principal's objective is to maximize profit subject to the participation (individual rationality) and incentive compatibility constraints.

The incentive compatibility concept, and principal agent analysis, are of relatively new origin. Principal agent analysis is mainly technical, quantitative, and mostly carried out by mathematical economists, who usually hold that there is an inevitable trade-off between equity and efficiency. To the extent that principal-agent analysis does not claim to be an ethical theory, it is mandatory that we evaluate the ethical implications of the principal-agent approach.

Dees (1988, p. 21) states that principal-agent analysis often "shifts our attention from problems of predicting complex social behaviour and developing social policy to a problem of egocentric strategy". He writes that principal-agent models do not go far beyond answering questions about what contractual arrangements will serve the various actors' interests. The models say nothing about what contract or joint solution is likely to result from bargaining (Cramton and Dees, 1988) between the various actors, and they go only a short way in evaluating what kind of contractual arrangements are to be preferred from an ethical point of view, and how we can ensure commitment to socially desirable arrangements *ex post*. Dees' (1988, p. 13) key argument is that principal-agent analysis "has framing effects that could blind less cautious users to important ethical dimensions of contractual arrangements". These are to the highest extent realistic dangers. Examples of biases that may ensue are that model users may: ignore the principal's obligations to the agent; develop excessive distrust and disrespect for agents; overlook ethical constraints, such as fairness; and miss solution possibilities that include ethical norms.

Our objective in this paper, by concentrating our attention on the individual agent, is to evaluate to what extent reputational considerations can be used to counter or align these potentially harmful framing effects or biases. It is clear that reputational considerations are relevant regarding all the four mentioned biases. Here, our aim is to use the reputation concept in its fullest and broadest sense, and see how much it is possible to get out of it. One reason that no other considerations than reputation are taken into account is that our focus is the individual agent. We have started out in this article by accepting only those constraints on self-interested behaviour that the individual agent agrees to by free will. This is in a sense a solid rock bottom from which we start out. The approach may thus possibly sound plausible even to the most "hard-headed street corner nihilist". Although we may desire that our practices should or ought to be intrinsically other-regarding, or justifiable according to what we consider as "our social womb", such aspects have no weight in our analysis except indirectly through reputational considerations.

Some philosophical support may prove fruitful for the approach. A plausible choice is Hobbes (1651), whose ideas underlie much of economic theory. Hobbes (1651, ch. X) defines "the Power of a Man…[as] his present means, to obtain some future apparent Good", and claims that man, instead of having sympathetic feelings, is self-interested, i.e. has will to power. The chain-store game (section 7) illustrated that a good reputation in early stages in the game proves fruitful in later stages. Hence a good reputation early in the game is a means to obtain increased utility later in the game. This implies that there is a sense in which we can substitute reputation for power in Hobbes' power definition. Hobbes (1651, ch. XI) "put[s] for a general inclination of all mankind, a perpetual and restless desire of Power after power, that ceaseth only in Death". The approach in this article is analogous, although employing more decent language. We say that our agent is self-interested, and has a desire for good reputation.

It may be remarked that our approach in one specific sense stands partly in contrast to Hobbes' theory. In defining power as a means to obtain some future apparent good Hobbes includes in this good inner psychological states of preferable character (e.g. giving a beggar money because it makes one feel good) as well as other features contributing to what one might define as a good. This article does not focus on inner psychological states, or other features contributing to some good, except in so far as these psychological states or other features contribute to a beneficial reputation. This illustrates that an alternative to considering our reputation concept as corresponding to Hobbes' power concept is to consider it as corresponding to Hobbes' value concept. Hobbes considers power as something one has for oneself, whereas value is something one has for others. He (Hobbes, 1651, ch. X) says that "the Value, or Worth of a man, is as of all other things, his Price; that is to say, so much as would be given for the use of his Power". Hobbes' value theory is subjective and conditional, so that "the buyer determines the Price", and "a learned and uncorrupt judge is much Worth in time of Peace; but not so much in War". This is also the case for our reputation concept.

To sum up, our agent is self-interested and has a desire for good reputation. Our objective is to evaluate how far this brings us in ethical analysis. We seek to pronounce explicitly where the real interests of the individual agent actually lie. Such elucidation and illustration may prove beneficial not only for the agent, but also possibly for others interacting with the agent, and for policy makers structuring the surroundings in which the agent functions. Further, illustration of reputational aspects is relevant to counter a possible belief that a main driving force in today's society, especially when economic considerations are at stake, is narrow, short-sighted self-interest.

Interaction between individual agent and surrounding structure

Standard economic textbooks typically describe actors with self-interest (with guile) and bounded rationality interacting in a world with uncertainty/ complexity and scarce resources. Self-interest and scarce resources are present in Gauthier's (1986) theory. Bounded rationality is a problem. Game theory assumes rational (i.e. each individual agent's decision-making behaviour is consistent with the maximization of subjective expected utility) and intelligent (i.e. each individual agent understands everything about the structure of the situation, including the fact that others are intelligent and rational) actors, although there are extensions to bounded rationality (Kreps, 1990, ch. 6). This assumption implies the requirement of a high capacity for processing information and preferences. Gauthier also seems to assume "fully" rational actors. Reasoning oneself forward to his Archimedean point requires at least some degree of information processing. Our approach assumes a "sufficient degree" of rationality.

Taking our five building blocks as granted, consider the individual agent. She has self-interest. She participates to the extent that doing so is individually rational. She reveals her true type to the extent that she has incentives to such behaviour. And she seeks a good reputation. We saw above that to ensure a role for reputation there must be some uncertainty regarding how the agent will act in the future, i.e. the agent must be unable to commit in one or the other way in advance. However, we have not illustrated to what extent some kind of reputation actually is relevant for the agent. And we have not illustrated what kind of reputation is beneficial. It is not possible to answer these two questions before we have said something about the surrounding structure in which our agent operates.

Consider the value systems of our agent, of the organization, and of the surroundings around the organization which we will call mankind. These do not typically overlap (Figure 2). An action done by our individual agent may be ethical or not according to the agent's own personal value system, according to

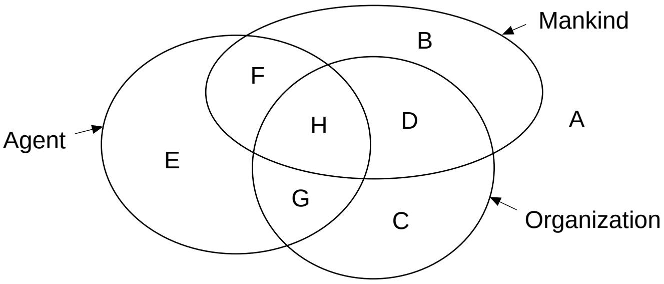


Figure 2.
Value system of agent, organization and mankind

the organization's value system as partly pronounced by the corporate culture, and/or according to mankind as partly pronounced by our cultural heritage, laws, customs, etc. Whence follows that the action can be evaluated along three dimensions (Table I). We aim to proceed towards what mankind collectively agrees on as ethically justifiable. We seek to increase overlap of value systems, especially increase area H. And we ask to what extent we can do this by the help of the reputation concept.

Whereas self-interest is something we have for ourselves, reputation is also judged by our surroundings. Hence briefly consider organization theory, after which we will work down to our individual agent again. A standard method in organization theory today is to conceptualize a 2 × 2 matrix labelled individual level, organizational level along the vertical axis, and internally directed behaviour, externally constrained behaviour horizontally. Most organizational theories can be placed somewhere within this two-dimensional dichotomy (Astley and Van de Ven, 1983). There is some indication to suggest a trend in organization

	A	B	C	D	E	F	G	H
Agent	0	0	0	0	1	1	1	1
Organization	0	0	1	1	0	0	1	1
Mankind	0	1	0	1	0	1	0	1

Notes:
A: Reprehensible action
B: May be due to inertia in our laws
C: Mafia organization having moralistic agents
D: Agent adhering to a specific non-traditional moralistic system
E: Criminal action
F: Idealistic organization
G: Mafia organization having talent in socializing agents
H: Ideal action

Table I.
Agent's action judged ethically OK (1) or not (0) according to the agent's, the organization's and mankind's value system

theory from the lower left to upper right in this matrix. To the extent that this is correct, established procedures, practices and culture at the organizational level, as well as how agents are socialized into the corporate culture, which values they are influenced to embody in their actions, and how they are trained in all respects, is of crucial importance. The surrounding structure and corporate culture may guide, influence and cement individual behaviour.

Kreps (1984), attempting to link reputation to economic theory, identifies corporate culture with a set of principles and the means by which these principles are communicated. The principles are preferably simple and consistent. An organization nurtures its reputation by establishing a culture which implicitly communicates these principles. The principles give hierarchical inferiors an idea *ex ante* how the organization will react to circumstances as they arise. Violation of the culture, by actors controlling and influencing cultural factors, generates direct negative externalities insofar as it weakens the overall reputation of the organization. A logical implication is that the culture will reign even when it is not "first best". A related implication is that an organization may find an interest in acting beyond narrow shortsighted self-interest.

Simplify and assume that the principal has the main control over the corporate culture, hence also over the organization's reputation. This is not always the case, e.g. stockholders may interfere, and agents necessarily influence as enacters. However, the principal has to some extent authority to control, direct and channel the use of resources.

Further simplify and assume that the principal's interests coincide with the organization's interests (which is also not always the case). Granted this, the principal must evaluate what kind of reputation to nurture for the organization. This evaluation is analogous to the reputation the individual agent finds it beneficial to nurture, but not quite. A realistic hypothesis is:

H1: what arises from a multitude is often more easily accounted for by determinate and known causes than by what depends on a few persons.

An organization consists of a multitude of actors and must be institutionalized in its surroundings in order to survive. Such considerations elucidate that an organization's reputation is crucial. A more speculative hypothesis is:

H2: An organization's reputation embodies to some extent features consonant with those values we honour and revere in today's society.

This hypothesis is not always fulfilled, as illustrated by the presence of mafia organizations whose reputations do not necessarily embody values universally embraced by mankind. Nevertheless, we hypothesize, without being too specific, that ethical concerns to some extent have to be represented in an organization's reputation. Further, based on Kreps (1984), we pose a third hypothesis:

H3: Features being present in an organization's reputation are to some extent represented in the organization's corporate culture, internal structure and the internal surroundings in general around the individual agent.

Simplifying, incorporate all these aspects around the individual agent into the one concept of corporate culture. Combining the last two hypotheses, consider the following hypothesis:

H4: To the extent that ethical concerns are relevant for an organization's reputation, these ethical concerns are also embodied, somehow, in the corporate culture.

Having said this much about organizational features, i.e. about the surroundings in which the agents (we assume several, although our main focus is the one single individual agent) act, we can return to the two questions we could not answer above. We asked to what extent reputation is relevant for the agent and what kind of reputation is beneficial. The answers to these questions lie implicitly in a verbalization of the corporate culture. Hence recapitulate and qualify what we assumed above about the individual agent. The individual agent has self-interest, she participates to the extent that doing so is individually rational, and she reveals her type to the extent that she has incentives to such behaviour. We can now add that our agent seeks reputation to the extent that reputational considerations are relevant, as implicitly, tacitly, etc. spelled out through a verbalization of the corporate culture. Further, our agent seeks that specific reputation which is beneficial. Information about what kind of reputation is beneficial also lies embodied, somehow, in the corporate culture. In other words, the corporate culture signals to the agents, openly, covertly, implicitly, tacitly and in an extensively high number of known and unknown ways, to what extent reputational considerations are relevant, as well as what kind of reputation is beneficial. On the basis of this, and consonant with our fourth hypothesis, consider the following hypothesis:

H5: To the extent that ethical concerns are embodied in the corporate culture, these ethical concerns are also relevant for our individual agent's reputation.

Combining *H2-5* gives a sixth hypothesis:

H6: Those ethical values we honour and revere in today's society are to some extent relevant for the kind of reputation for which our individual agent finds it beneficial to strive.

What we have possibly illustrated with this link from mankind to our individual agent is that what the former as such revere today may be relevant for the reputation of the individual agent. Consider another hypothesis:

H7: Increasing the role of reputations in wider contexts may ensure that these reputations to a higher extent embody those values which mankind reveres in today's society than if reputations are rendered to have a role only in the narrow surroundings around the individual agent.

One idea underlying this hypothesis is that if reputations have significance only in the narrow surroundings around the individual agent, this agent, or a small coalition of agents, may find it more beneficial to undertake (covert) free-riding with potential immediate gain, rather than taking long-term reputational considerations into account. To the extent that this hypothesis is fulfilled, our objective is to make the link from mankind to our individual agent strong, and to make mechanisms, procedures and underlying assumptions explicitly pronounced. Another hypothesis is:

H8: To the extent that the link from mankind to our individual agent is strong, we are enabled to assign more "ethical content" to our basic framework than if the agent was rendered to seek a reputation in her more narrow surroundings.

Hence, analyse features present on the first level above individual agent, i.e. features of the corporate culture. We thereafter provide the principal with advice in the forming of corporate culture. This leads to an immediate problem present in all ethical analysis: the corporate culture is seldom explicitly pronounced. As observed above, Kreps (1984) believes that communication of the mentioned principles is crucial to establish a reputation for the organization. And, as observed, corporate culture (Kreps, 1984, p. 42) "gives hierarchical inferiors an idea *ex ante* how the organization will 'react' to circumstances as they arise". I believe that all ethical analysis in organizations benefit from a clear verbalization of the crucial features of the corporate culture, i.e. by making mechanisms explicitly pronounced. It is clear that the corporate culture may implicitly (tacitly, etc.) encourage or discourage "ethical" behaviour. The problem is that there are counterforces against explicit verbalization. The agents must be "trained". Not everything can be foreseen *ex ante*. And the presence of hidden action/information, combined with a perhaps imperfect corporate culture, may induce agents, alone or in coalitions, in "unfavourable" directions.

Consider the advice to the principal regarding what kind of corporate culture to establish, granted that she has some kind of control in this regard. As we observed in the principal agent models in the sections on incentive compatibility, and principle-agent analysis and ethics, it is typically the case with uncertainty and asymmetric information that the agent is better informed about her own effort level and type than the principal. Thus, the principal has interest in ensuring a role for reputations (to the extent that the agents cannot be committed in one way or the other), as well as an interest in making reputations publicly known. The first two pieces of advice are related to this. The first piece of advice follows from Wilson's (1985) observation that a key insight in ensuring a role for reputations is to provide for strong intertemporal linkages along a sequence of otherwise independent situations. The second piece of advice involves establishing a culture where the market for reputational information has a high degree of efficiency. The third piece of advice to the principal is to establish a culture which implicitly gives incentives to the agents to nurture that specific reputation which is beneficial for the reputation of the organization. Hence, the principal, in establishing corporate culture, is influenced by external factors regarding what kind of reputation to nurture for the organization as such. However, she is also influenced by internal factors in the organization regarding how to figure out and lay the foundation for a culture which implicitly induces the agents to strive for that specific reputation which is beneficial for the organization.

To recapitulate, the principal seeks to maximize profit by nurturing a beneficial reputation for the organization, which is done partly through establishing corporate culture. The corporate culture implicitly ensures a role for reputations for agents by providing for strong intertemporal linkages along sequences of otherwise independent situations, ensures an efficient market for reputational information and embodies values revered by mankind. Proceeding from this, we observed that to the extent that ethical values are somehow embodied in the corporate culture, and to the extent that mechanisms, procedures, and underlying assumptions in general are verbalized and explicitly pronounced, our individual agent has an interest in striving for a reputation which to some extent embodies values revered by mankind or contemporary society in general. Such values may be honesty, fairness, trustworthiness, altruism, justice, other-regarding behaviour, respect for the autonomy of others, a sense of public duty, and perhaps equality, objectivity, impartiality, intellectual integrity, etc.

The real interests of the individual agent in the long run

To return to the individual agent, the key object under scrutiny in this article. The individual agent is to some extent rendered to take the corporate culture of the organization as a given state. This state may be perceived as controlled by the principal, or as resulting from historical evolution, influenced by forces and counterforces. The agent has self-interest, which involves seeking reputation. She participates to the extent that doing so is individually rational, and she reveals her type to the extent that she has incentives to such behaviour. Further, she asks two crucial questions:

(1) Does reputation play a role here?

(2) What kind of reputation is beneficial?

Our agent implicitly, through her choice of strategy, answers these questions by making efforts to establish her own reputation, roughly proportional to how strong the intertemporal linkages along sequences of otherwise independent situations might be, and roughly proportional to how efficient the internal market for reputational reputation is. She finally seeks to nurture that specific reputation which, through the corporate culture, is spelled out as beneficial for the reputation of the organization.

Consider one possible objection to our framework. It might be argued that the premisses from which it starts out are "unethical", since there is nothing here except pure self-interest. Is this what we ought to teach the next generation? The response to this is that our and the next generation have much to learn, about basic honourable values, about how to avoid narrow short-sighted self-interest, etc. Whether a theory is ethical or unethical depends on judgement. Philosophers disagree regarding judging theories as ethical or unethical. Adam Smith focuses on moral sentiments, Hume on sympathy and motives, Christianity on universal love, Kant on a universal law, Teleologists on consequences. The Stoic says that to be conscious of one's virtue is happiness. The Epicurean says that to be conscious of one's maxims as leading to happiness is virtue, etc.

Philosophers go beneath those values which are dominant in a given society. To some extent philosophers are removed from the current spirit diffused throughout the people at a given time and place. This may explain the fact that even though philosophers disagree, in some sense there seems to be remarkable agreement regarding what is ethical/unethical. As we mentioned above, we seem to agree that ethics has to do with such concepts as honesty, fairness, trustworthiness, altruism, justice, other-regarding behaviour, etc. If we ask the common man to justify what it specifically is that makes these concepts ethical or honourable, he may start by saying that the relevance and weight of the various concepts may depend on time and place. However, sooner or later our discussion will lead us to conclude that we are dealing with belief, with insufficient justification. People's beliefs do not entirely overlap. Differences in culture, socialization, habit, interest, ability, give rise to a more or less explicitly pronounced hierarchy of moral values from which to choose actions. An intrinsic problem in ethical analysis is that this hierarchy of values is such that we can to some extent expect a difference between what values a person espouses or publicly assents to, and what values are revealed through action, hidden or not. A main objective is to render mechanisms, procedures, and underlying assumptions explicitly pronounced, thereby possibly increasing the actual adherence to publicly "legitimate" values, and thereby providing agents with incentives to strive for a reputation embodying higher weight and relevance of ethical values.

Returning to our framework, one might try to argue that what is worthy of critique in our approach is that our individual agent seeks beneficial reputation for this reputation's own sake, whereas what our agent "ought" to do is to seek something else, e.g. something intrinsically "ethical". Scrutinize the root of this critique. In Christian thought one is advised first and foremost to seek the kingdom of God, with a promise of then subsequently benefiting also in other respects. A set of values is considered dominant in today's society. The content of this set is influenced by a high number of different sources, among them Christianity. We may feel that there is something in this value-set saying that one “should” not seek reputation for its own sake, but something else. We are occasionally criticized for living in a commercial society. We may have a rather vague notion of what this “something else” should be, e.g. God, mankind, our country, democracy, human rights, the poor, the sick, our neighbour, etc. Several further evaluations might be made about this “something else”. However, observe that if seeking this “something else” is considered ethical, then getting a reputation for seeking “something else” may be a natural by-product. In our model our agent seeks to nurture that specific reputation which, through the corporate culture, is spelled out as beneficial. Thus, if ethics or this “something else” is spelled out as beneficial, focusing on ethics or this “something else” may contribute to a beneficial reputation. In other words, if the corporate culture is of the indicated kind, our agent will find it beneficial to have a reputation for focusing on ethics or this so-called “something else”. The difference, then, between seeking reputation for its own sake, or seeking “something else” with reputation as a potential by-product, has to do with the motives for human behaviour, not with the consequences following from this human behaviour. Economic theory has traditionally a stronger focus on consequences than on motives. However, observe that ethical motives may have indirect influence in our framework, by contributing to an agent’s reputation. Hence, if we skip the cumbersome part regarding whether our agent seeks reputation for its own sake, or whether she seeks “something else” with reputation as a potential by-product, our framework, as a descriptive explanation of human ethical behaviour, may work in both cases. And what we may potentially obtain is the foundation of an ethical theory which is compatible with today’s economic theory, a compatibility seemingly lacking in many of today’s ethical theories. We observed in the section on “principal-agent analysis and ethics” that our framework is compatible with Hobbes’ economic and political theory. But inherent and implicit in the reputation concept lies adherence to those ethical values we revere in today’s society.

Although accepting our theory as potentially ethical, our objector may argue that reputation is often a weak and ineffective reason for an agent to constrain behaviour and curtail abuse. This is an objection worthy of serious scrutiny. The actor’s cost/benefit analysis may involve building, maintaining and “milking” her reputation in a way that disregards ethical concerns. I have four comments on this, the first three related to the advice we gave to the principal in the last section regarding what kind of structure and corporate culture to establish. First, to ensure a role for reputations one must provide for strong intertemporal linkages along a sequence of otherwise independent situations. Second, an efficient market for reputational information renders reputational information crucial. Third, to the extent that ethical values are important in contemporary society a reputation for adherence to these values is important. Finally, an agent will in her cost/benefit analysis have an interest in undertaking extensive and thorough research into what long-term reputational considerations actually are relevant.

Illustrate with an example features related to these first two comments, intertemporal linkages and efficiency of market for reputational information. Assume a principal and an agent in a repeated prisoner's dilemma with honest and deceitful actions where mutual honesty gives pay-off (2, 2), mutual deception (1, 1), free-riding 3, whereas being exploited gives 0. Detection of deception implies mutual deception for the remainder of the game. Assume discount factor δ, and that deception is detected with probability q. Mutual honesty yields 2 in every period with an expected payoff of $2/(1 - \delta)$. Unilateral deception in the first period gives 3, and then 1 or 2 thereafter; i.e. detection of deception in the first period yields expected payoff of $1/(1 - \delta)$ with probability q, and $2/(1 - \delta)$ with probability $1 - q$, considered from period 2 and thereafter. If the actor contemplates the two scenarios of co-operating forever on the one hand, or deceiving in the first period and then co-operating forever thereafter on the other hand, he will choose the former if

$$\frac{2}{1-\delta} \geq 3+\delta\left[q\frac{1}{1-\delta}+(1-q)\frac{2}{1-\delta}\right], \tag{14}$$

i.e. if $\delta \geq 1((1 + q)$, which is exponentially declining in q. Hence, if there is a 100 per cent chance of detection of deception ($q = 1$), it is rational to act honestly if the discount factor is 1/2 or higher. With a 50 per cent chance of detection a discount factor of 2/3 is necessary to rationalize honesty. If the actor, on the other hand, contemplates the two scenarios of co-operating forever on the one hand, or deceiving for n periods and then co-operating forever thereafter, starting in period $n + 1$ on the other hand, he will choose the former if

$$\begin{aligned}\frac{2}{1-\delta} \quad &\geq \quad 3+q\frac{\delta}{1-\delta}+(1-q)3\delta+(1-q)q\delta^2+(1-q)^2 3\delta^2\\ &+\ldots+(1-q)^{n-1}q\delta^n+(1-q)^n 3\delta^n+(1-q)^{n+1}2\delta^{n+1}\\ &+(1-q)^{n+1}2\delta^{n+1}+\ldots\\ &=3+\frac{\delta}{1-(1-q)\delta}\left[\frac{q}{1-\delta}+3(1-q)\left[1-(1-q)^n\delta^n\right]\right.\\ &\left.+2(1-q)^{n+1}\delta^n\right]. \end{aligned} \tag{15}$$

Equation (15) is most appropriately solved numerically. Whether it is more or less strict than equation (14) depends on q and n and, more generally, on the pay offs 2, 1, 3 and 0. Since the objective of this section is to illustrate the reasoning involved when the individual agent evaluates her real interests in the long run, curves and graphs are left out as these are more appropriate for a more specific analysis. What can be said in general is that the more unlikely detection of deception becomes the higher emphasis on the future is necessary to rationalize acting honestly. Features of this kind are crucial in the Folk Theorem (Fudenberg and Maskin, 1986), which states that any individually rational pay off vector in a one-shot game of complete information can arise in a perfect equilibrium of the corresponding infinitely repeated game if the players discount future sufficiently little (δ is high). Thus, regarding our two first recommendations to the principal, to the extent that she nurtures honest behaviour among the agents, she should establish a corporate culture which enlarges the shadow of the future (Axelrod, 1984), i.e. renders something distant (future) the immediate interest of the agent. It is beyond the scope of this paper to investigate how the principal should do this. Axelrod (1984, p. 126) recommends making interactions more durable and frequent.

Our third recommendation to the principal was to nurture the organization's reputation by embodying ethical values in the corporate culture to the extent that such values are relevant in contemporary society. Hence reformulate our last example to see to what extent a reputation argument for the principal or the organization can be used to counter the first possible framing effect or bias we mentioned in the section on principal analysis and ethics; that principal-agent model users may tend to ignore the principal's obligations to the agent. In our repeated prisoner's dilemma both the principal and the agent are assumed equally at risk. However, principal-agent analysis usually assumes that the agent is better informed about her own type than the principal. This places the principal in a more risky position, which our agent may exploit. Taken to its most extreme, assume that the principal is enduring, and meets a new agent in each new period of the game. It is thus rational for the agent to deceive. However, the principal may avoid this by structuring the corporate culture so that the agent acts first. If the agent moves first in a repeated prisoner's dilemma, the principal can condition her behaviour on the agent's action. The agent cannot develop a reputation, but the principal can, and will have an interest in doing so to the extent shown in the previous example. That is, to the extent that the future is sufficiently important to the principal (δ is high), and to the extent that agents can detect deception by the principal (q is high), the principal will have an interest in acting honestly, i.e. establishing a reputation for honest behaviour. This interest for the principal in acting honestly is thus derived from internal factors in the organization. The principal may also have an interest in a reputation for honest behaviour owing to external factors. The force of this last interest depends on the extent to which such a reputation is honoured by dominant interests in the surroundings around the organization. Our agent, knowing that she is at risk by having to move first in the repeated prisoner's dilemma, and granted that the principal acts honestly if she does, will have an interest in acting honestly, that is, she will act honestly as long as the principal has never deceived an agent in any period since the initiation of the sequential game.

Regarding the fourth comment above, return to the individual agent, who is rendered to take the current state of the corporate culture as given. I will elaborate on this last crucial point. An inherent problem we can work to reduce but never entirely avoid, is the presence of hidden action and hidden information. Machiavelli (1532) advised his Prince to appear to be honest. There are gains from appearing to deserve a good reputation, e.g. a reputation for adhering to values considered ethical in today's society. However, there are also gains from acting contrary to these ethical values (e.g. free-riding, exploitation, cheating). Label this last type of behaviour deceptive. From this follows that there are gains from being able to preserve one's reputation when also deceiving, e.g. cheating without being detected. As we have seen, deception typically gives immediate gain whereas reputation involves long-term considerations. Because of this we have advised our principal to construct a corporate culture "enlarging the shadow of the future", inducing the agent to think long term. This leads to a crucial point. Our aim is to argue that our agent actually has a real interest in thinking long-term for other reasons than those that might be indicated by a narrow interpretation of the corporate culture. Given the state of the corporate culture, is our agent a good utility-maximizer in the long run? To what extent does our agent attach weight to relevant occurrences in the distant future? One theory of human nature (Frank, 1988, p. 78) is that "the attractiveness of a reward is inversely proportional to its delay". Of course, the discount factor accounts for some of this. However, we argue that our agent is likely to have an interest in discounting future to a lesser degree than a narrow economic interpretation would imply. Assessments regarding problems like this are of course complicated to make, owing to complexity and other factors. We do not know the future, and we may be in doubt regarding what reputation is beneficial. Nevertheless, our agent continuously evaluates the gain per time unit of her reputation against the gain per time unit of deceptive behaviour. To the extent that our agent is influenced by the psychological theory mentioned above (Frank, 1988), she will be more likely to deceive. To the extent that deception is done without talent, the risk of detection increases, implying potential negative effect on reputation. Reputation is an observable characteristic. We are therefore naturally led to evaluate (Frank, 1988) the relation between the reputation and the true (unobservable) character.

An agent's character depends on previous known and unknown history. We might venture the hypothesis that a high degree of deception in previous history may influence character negatively, hence subsequently influencing reputation. For example, acting virtuously may possibly lead one to become virtuous, and hence establish a virtuous reputation. Conversely, acting viciously may perhaps (but not necessarily) lead one to become vicious, unless one is able to distance oneself from one's own action. For example, one may attempt to legitimate some types of behaviour standing in contrast to one's personal value system by saying that one is acting as a representative for an institution or for other interests where other values are crucial, thereby in a sense detaching oneself from one's actions. A careful analysis of implications of virtuous/vicious behaviour on character and subsequently on reputation involves venturing into the psychology field, which is outside the objective of this article. Frank (1988, p. 83) uses the term impulse control. Agents are likely to have less than 100 per cent impulse control.

We may hypothesize that being predisposed to act ethically may indirectly lead to a good reputation, since one avoids falling victim to imperfect impulse control, e.g. being detected in a dishonest action. An important question is whether the agent, in deceiving, would have allowed other agents in the same situation to deceive. If not, our agent to some extent detaches herself from a basic universalistic principle in our culture, a detachment we may hypothesize (this is not necessarily true) will have negative influence on character, and hence potentially on reputation. Nevertheless, to the extent that our agent has perfect impulse control and can perfectly foresee whether detection is possible, there will be no link between reputation and character. To the extent that our agent does not deceive we can say that her reputation gives a true picture of her character. To the extent that she deceives without being detected she gains by this behaviour and by having her reputation intact. The crucial question is whether our agent is able to deceive without being detected in the long run. To the extent that deception without detection is possible, our reputation framework, as an ethical theory for the individual agent, may have bad ethical implications, suggesting a need to search for remedies in the surroundings around the individual agent, which may be a topic for a future article. Taking the surroundings as given, focusing on self-interest, individual rationality, sequential rationality, incentive compatibility, and reputation is a crucial starting point for the individual agent.

References and further reading

Astley, W. and Van de Ven, A.H. (1983), "Central perspectives and debates in organization theory", *Administrative Science Quarterly,* Vol. 28, pp. 245-73.

Axelrod, R. (1984), *The Evolution of Co-operation,* Basic Books, New York, NY.

Cramton, P.C. (1988), "Strategic delay in bargaining with two-sided uncertainty", manuscript, Yale University, New Haven, CT.

Cramton, P.C. and Dees, J.G. (1988), "Deception in negotiation: a study of the relationship between self-interest and ethics", manuscript, Yale University, New Haven, CT.

Dees, J.G. (1988), "Principals, agents, and ethics", manuscript, Harvard University, Cambridge, MA.

Elster, J. (1989), *The Cement of Society,* Cambridge University Press, Cambridge.

Frank, R.H. (1988), *Passions within Reason,* W.W. Norton & Co, New York, NY.

Fudenberg, D. and Maskin, E. (1986), "The folk theorem in two-person repeated games with discounting and incomplete information", *Econometrica,* Vol. 54, pp. 533-54.

Gauthier, D.P. (1986), *Morals by Agreement,* Oxford University Press, Oxford.

Harsanyi, J.C. (1967-68), "Games with incomplete information played by 'Bayesian' players, I-III", *Management Science,* Vol. 14, pp. 159-83; 320-34; 486-501.

Hobbes, T. (1651), *Leviathan,* J.M. Dent & Sons, 1973 edition.

Hume, D. (1740), *A Treatise of Human Nature,* Selby-Bigge, L.A. (Ed.), Clarendon, Oxford, 1978 edition.

Hume, D. (1751), *An Enquiry Concerning the Principles of Morals,* Oxford University Press, Oxford, 1975 edition.

Kreps, D.M. (1984), "Corporate culture and economic theory", manuscript, Stanford.

Kreps, D.M. (1988), *Notes on the Theory of Choice,* Westview Press, Boulder, CO.

Kreps, D.M. (1990), *Game Theory and Economic Modelling*, Clarendon, Oxford.

Kreps, D.M. and Wilson, R. (1982a), "Sequential equilibria", *Econometrica,* Vol. 50, pp. 863-94.

Kreps, D.M. and Wilson, R. (1982b), "Reputation and imperfect information", *JET,* Vol. 27, pp. 253-79.

Machiavelli, N. (1532), *The Prince.*

Rawls, J. (1971), *A Theory of Justice,* Harvard University Press, Cambridge, MA.

Schelling, T.C. (1960), *The Strategy of Conflict,* Harvard University Press, Cambridge, MA.

Schmidtz, D. (1989), "The limits of government: an essay on the public goods argument, Yale University", manuscript, Yale University

Taylor, M. (1987), *The Possibility of Co-operation,* Cambridge University Press, Cambridge.

Wilson, R. (1985), "Reputations in games and markets", in Roth, A.E. (Ed.), *Game-Theoretic Models of Bargaining,* Cambridge University Press, Cambridge, pp. 27-62.