Ethical Approaches and Autonomous Systems

T.J.M. Bench-Capon
Department of Computer Science
University of Liverpool
Liverpool
UK.
e-mail tbc@liverpool.uk.ac

Abstract

In this paper we consider how the three main approaches to ethics – deontology, consequentialism and virtue ethics – relate to the implementation of ethical agents. We provide a description of each approach and how agents might be implemented by designers following the different approaches. Although there are numerous examples of agents implemented within the consequentialist and deontological approaches, this is not so for virtue ethics. We therefore propose a novel means of implementing agents within the virtue ethics approach. It is seen that each approach has its own particular strengths and weaknesses when considered as the basis for implementing ethical agents, and that the different approaches are appropriate to different kinds of system.

1. Introduction

Autonomous machines are already a significant feature of modern society and are becoming ever more pervasive. Although they bring undoubted benefits, this is accompanied by concerns. It is important that such machines function as a benign part of our society, and this means that they should behave in an ethical manner. Conventional machines are required to meet safety standards (e.g. cars are subject to regular checks to ensure that they are not dangerous), but autonomous machines additionally need to come with some assurance that their decisions are not harmful. In this paper we will consider how various approaches to ethics play out when applied to autonomous systems. We will consider three types of ethical theory: consequentialist approaches [55], deontological approaches [2] and virtue ethics approaches [36]. The intention is certainly not to argue that one approach is better than another as an account of human ethics, but rather to explore the effects of adopting the various approaches when implementing ethical agents.

Our notion of an ethical agent is *an agent that behaves in a manner which would be considered ethical in a human being*, This may be considered to be in the spirit of "weak" AI, and adapts Minky's definition of AI which relates to "activity produced by a machine that would have been considered intelligent if produced by a human being" [46]. The agents in our discussion will not be expected to model or perform ethical

reasoning for themselves, but merely to behave ethically. The approaches to ethics we discuss will belong to the system designers, not the agents, and we will be exploring the consequences of the designers adopting a particular approach to ethics on the agents they implement. This may result in the implemented agents representing a rather thin version of the approach they embody: for example, deontological agents will be rule followers, but will not discover, or be the source of, the rules they follow, nor will they follow them for the sake of the moral law. Both of these features are important parts of the deontological position as proposed in e.g. [37], but it will be the designer who identifies the rules and who, for the sake of the moral law, builds agents that will follow them.

By adopting this relatively weak notion of what it means for an agent be ethical, we are able to exclude from consideration anything dependent on mental states, or motivation, and questions such as whether the ethics developed by an agent, whose needs and form of life greatly differ from those of humans, would be acceptable to humans. We believe that using this notion is not a problem: on the contrary ethical agents are needed *now*, and while weakly ethical agents are currently feasible, strongly ethical agents currently lie in the, perhaps distant, future.

Currently agents built according to the consequentialist and deontological approaches are not uncommon, but there are few if any available following the virtue ethics approach. For this reason I will propose in this paper a means of enabling an agent to behave in accordance with virtue ethics, based on the approach to practical reasoning developed in [6] and subsequent papers including [7], [8] and [9]. This will represent one of the novel contributions of the paper. The other is that it provides an explicit discussion and comparison of the effects of designers adopting the various approaches on the systems they implement. Although many systems do embody ethical approaches, these approaches are often taken as given and there is rarely explicit discussion or justification of the design choice of adopting a particular approach to ethics. It is hoped that the discussion here will help to inform the choice of ethical approach for those wishing to build ethical agents in future.

2. Ethical Approaches

Broadly speaking, current ethical theories can be divided into three types: consequentialist approaches [55], deontological approaches [2] and virtue ethics approaches [36]. In this section we will briefly describe each of them.

2.1. Consequentialism

This approach holds that the normative properties of an act depend only on the consequences of that act. Thus whether an act is considered morally right can be determined by examining the consequences of that act: either of the act itself (*act utilitarianism*, associated with Jeremy Bentham, [20]) or of the existence a general rule requiring acts of that kind (*rule utilitarianism*, often associated with John Stuart Mill [45]). This gives rise to the question of how the consequences are assessed. Both Bentham and

Mill said it should be in terms of "the greatest happiness of the greatest number", although Mill took a more refined view of what should count as happiness[1]. However, there are a number of problems associated with this notion, and many varieties of pluralistic consequentialism have been suggested as alternatives to hedonistic utilitarianism (see [55]). Equally there are problems associated with which consequences need to be considered: the consequences of an action are often not determinate, and may ramify far into an unforeseeable future. However, criticisms based on the impossible requirement to calculate all consequences of each act for every person, are based on a misunderstanding. The principle is not intended as a decision procedure but as a criterion for judging actions: Bentham wrote "It is not to be expected that this process [the hedonic calculus] should be strictly pursued previously to every moral judgment." [20]. Despite the difficulty of determining all the actual consequences of an act (let alone a rule), there are usually good reasons to believe that an action (or rule) will increase or decrease general utility, and such reasons should guide the choices of a consequentialist agent. This has led some to distinguish between *actualism* and *probabilism* (e.g. [32] and [39]). On the latter view, actions are judged not against actual consequences, but against the expected consequences, given the probability of the various possible futures. Given our notion of a weakly ethical agent, the agent itself will be supplied with a utility function, and will choose actions that attempt to maximise it. The choice of the utility function, and the manner in which consequences are calculated will be the responsibility of the designer.

2.2. Deontological Ethics

The key element of deontological ethics is that the moral worth of an action is judged by its conformity to a set of rules, irrespective of its consequences. One example is the ethical philosophy of Kant [37]; a more contemporary example is the work of Korsgaard [38]. The approach requires that it is possible to find a suitable, objective, set of moral principles. At the heart of Kant's work is the the categorical imperative, the concept that one must act only according to that precept which he or she would will to become a universal law, so that the rules themselves are grounded in reason alone. Another way of generating the norms is offered by Rawls' *Theory of Justice* [51], in which the norms correspond to principles acceptable under a suitably described social contract. The principles advocated by Scanlon in [52] are those which no one could "reasonably reject". Divine commands can offer another source of norms to believers. Given our weak notion of an ethical agent which requires only ethical behaviour, the rules to be followed will be chosen by the designer, and the agent itself will be a mere rule follower. Thus the agent itself will embody only a very unsophisticated part of deontology: any sophistication resides in the designer who develops the set of rules which the agent will follow.

Problems of deontological ethics include the possibility of normative conflicts (a problem much addressed in AI and Law, e.g. [48]) and the fact that obeying a rule

[1] Mill wrote in [45] "It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied." Bentham, in contrast stated "If the quantity of pleasure be the same, pushpin is as good as poetry."

can have clearly undesirable consequences. Many are the situations when it can be considered wrong to obey a law of the land, and it is not hard to envisage situations where there are arguments that it would be wrong to obey a moral law also. Some of this may be handled by exceptions (which may be seen as modifications which legitimise violation of the general rule in certain prescribed circumstances) to the rules. Exceptions abound in law and their representation has been much discussed in AI and Law: for example exceptions to the US 4th Amendment in, e.g. [14] and [19]. Envisaging all exceptions, however, is as impossible as foreseeing all the consequences of an action. For this reason laws are often couched in vague terms ("reasonable cause" and the like) so that particular cases can be decided by the courts in the light of their particular facts. This will mean that whether the rule has been followed or not may require interpretation.

2.3. Virtue Ethics

Virtue ethics is the oldest of the three approaches and can be traced back to Plato, Aristotle and Confucius. Its modern re-emergence can be found in [4]. The basic idea here is that morally good actions will exemplify virtues and morally bad actions will exemplify vices. Traditional virtue ethics are based on the notion of *Eudaimonia* [50], usually translated as *happiness* or *flourishing*. The idea here is that virtues are traits which contribute to or are a constituent of *Eudaimonia*. Alternatives are; agent based virtue ethics, which "understands rightness in terms of good motivations and wrongness in terms of the having of bad (or insufficiently good) motives" [56], target centered virtue ethics [58], which holds that we already have a passable idea of which traits are virtues and what they involve, and Platonist virtue ethics, inspired by the discussion of virtues in Plato's dialogues [26]. There is thus a wide variety of flavours of virtue ethics, but all of them have in common the idea that an important characteristic of virtue ethics is that it recognizes diverse kinds of moral reasons for action, and has some method (corresponding to *phronesis* (practical wisdom) in ancient Greek philosophy) for considering these things when deciding how to act. Because the are few exemplars of implementations using the virtue ethics approach in agent systems, we will provide our own way of implementing a version of virtue based ethics in an agent system, based on value-based practical reasoning [6], which shows how an agent can choose an action in the face of competing concerns. The various varieties of virtue ethics do, of course, have a lot more to them, but again, given that we are considering weakly ethical agents, these considerations and the particular conception of virtue will belong to the designers, who will implement their agents to behave in accordance with their notions of virtuous behaviour, through the provision of a procedure for evaluating competing values.

3. Scenario

To explore these approaches we will use a scenario modelled as state transition diagrams. This scenario was introduced in [21] and used to explore the emergence and representation of norms in [18]. In this scenario the agents are capable of enacting the

roles in two morally didactic stories: the fable of *The Ant and the Grasshopper*[2] and a version of the biblical parable of the *Prodigal Son*[3]. In those stories in summer an agent can choose to play or to work. Work will build up a stock of food. When winter comes, if an agent played rather than worked it will have no food, and will have to ask a worker agent for food. If the worker does not give food it will die. Working produces a surplus, and so the worker could give food and still have enough for itself. Food does not last through the next summer, and so at harvest and the end of winter (carnival) there is feasting for those who have a surplus. In the fable, the ant works, while the grasshopper plays. When winter comes that ant refuses to share his food because it was the grasshopper's choice to be without food. In the parable, however, when the father works and the son plays, the father forgives the son and gives him food.

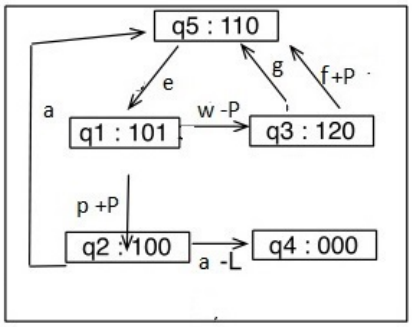

Figure 1: State transition diagram for each agent in the scenario. P is pleasure, L is life, w is work, p is play, a is ask, g is give, e is eat and f is feast.

Individual agents in this scenario can be represented by the state transition diagram[4] shown in Figure 1. States represent whether the agent is alive (1) or dead (0), whether it has no food (0), enough food (1) or surplus food (2), and whether it summer (1) or winter (0). In q1 the agent can choose to work (demoting pleasure) or to play (promoting pleasure). Choosing play means that in the next state asking for food is the only option, but whether asking will result in getting food or death is not within the agent's control. Working moves to q3, where the agent may choose to feast (promoting pleasure) or to give the surplus food away. This will promote the life of the other agent, but serves no value for the agent itself. In the fable, the ant chooses to feast, in the parable the father chooses to give.

Figure 2 shows the state transition for the community of agents as a whole. On this view, the actions of individuals cannot be distinguished: actions from q1 are all work, all play or some work and some play. This view is useful when determining what is

[2]One of Aesop's Fables, numbered 373 in the Perry Index.

[3]Luke 15:11-32

[4]The particular form of state transition diagram we use is an Action Based Alternating Transition System with values [6]. Where there is more than one agent, the transitions represent the concaternation of the (independently made) choices of the agents. The transitions are labelled both with actions and with the social values promoted or demoted. A formal definition can be found in the Appendix.

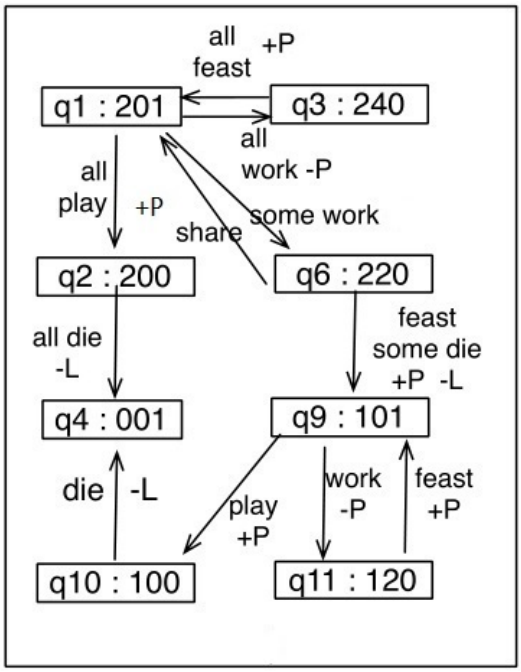

Figure 2: State transition diagram for community of agents in the scenario. P is pleasure and L is life.

best for the community: for example if one is thinking about what should become the law for that community.

Figure 3 shows a final view, showing the interaction between two agents. In this scenario one can see what actions by the other lead to particular outcomes, so that this can be taken into account when choosing ones own action. This is in contrast with the individual perspective of Figure 1, when whether the transition for q2 goes to q4 or q5 has no explanation, and Figure 2 which does not distinguish individuals. The labels also indicate which agents perform which actions, and which agent the values are promoted in respect of.

Agents in the scenario may have two choices to make.

MQ1 In summer both agents have a choice of whether to work or play.

MQ2 In winter an agent who has worked in the summer may have a choice as to whether to give some of its food away to an agent who played.

In the following sections we will consider these questions from each of the three ethical perspectives.

4. Consequentialism in the Scenario

On the consequentialist approach the ethically correct decision should be determinable simply by considering the consequences of the alternative actions for the so-

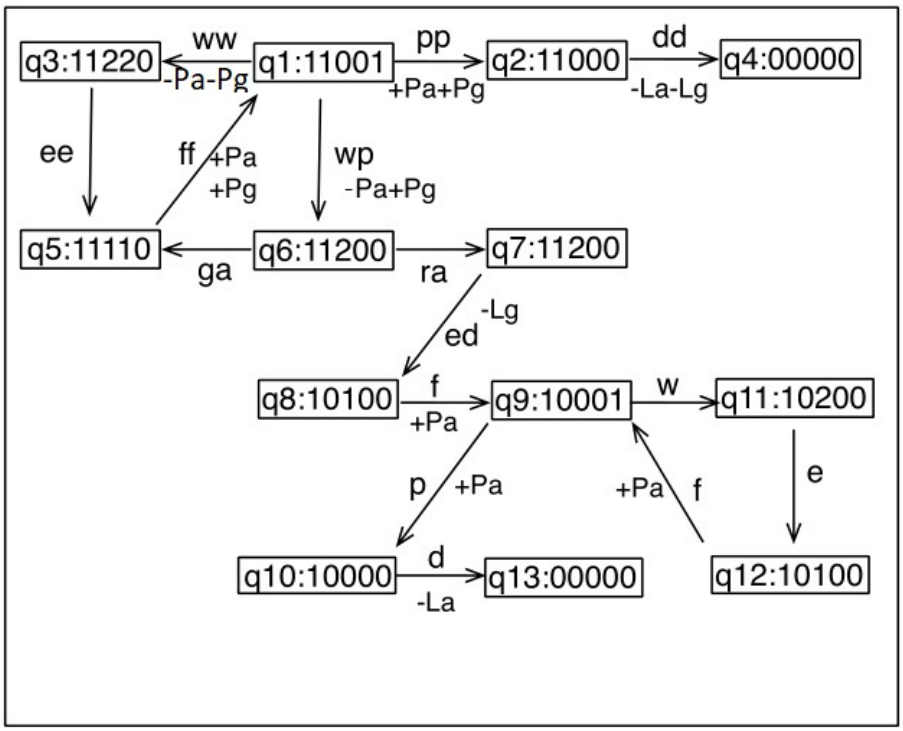

Figure 3: State transition diagram showing interactions within the community of agents in the scenario. Values are distinguished according to which agent they are promoted in respect of: e.g. Pa is pleasure for the ant, and Lg is the life of the grasshopper. Joint actions are writen as e.g. wp, which means that one works and one plays. In some states there is only one agent, and so only one action.

ciety as a whole.

4.1. Consequentialism and MQ1

MQ1, the question of whether to work or play, arises in q1. From Figure 1 we can see that the immediate consequences of play are pleasure, and that of work, demotion of pleasure. From a hedonistic utilitarian perspective the choice would appear to be simple: play. The problem is that play takes the agent to q2, where there is a possibility of death. If one then refers to Figure 2, it will be apparent that this possibility becomes a certainty if the other agent has made the same choice. Figure 3 reveals that even if the other agent has worked, survival depends on the worker sacrificing its own pleasure to support the player.

When making its choice, the agent must consider what others will do. Thus even if an agent thought that playing and reaching q6 in Figure 3 was better than working to reach q3, the risk of reaching the undesirable q2 has to be recognised. A prudent agent will not take this risk, so it will have to work unless it has reason to suppose that others will work and give away their surplus food rather than feast, or it has a sufficient

preference for pleasure over life. There is nothing necessarily immoral about such a preference. In Greek mythology Achilles was offered the choice between gaining glory and dying young or living a long life in obscurity. Achilles chose the former, and no one thought the worse of him for it. More recently James Dean's biography was entitled *Live Fast Die Young*, without implying criticism of Dean[5]. However, it is not the choice that everyone would make, and so allowing it would either render our ethics relativistic to individual preferences, or require that we regarded MQ1 as morally neutral.

In both our stories, however, there is strong implication that the answer to MQ1 should be to work. In the fable the suggestion is that the grasshopper deserves punishment for the summer of idleness, and in the parable the son is given food, but only because the father forgives him, implying disapproval of the aestival activity. How then can we provide a more objective basis to the preference for life over work?

One approach would be to appeal to the consequences for the community as shown in Figure 2. Here is clear that the community will cease to be if everyone plays, and one might consider sustainability to be the important concern of the community, and if an agent is permitted to play that is no longer possible. The only way an agent can ensure that the bad consequences do not come about is to work itself. While this might be found persuasive by some, there seems no knock down argument for favouring the interests of the community over one's own. Moreover on a strictly hedonistic utility principle, net pleasure only comes from play, and so perhaps the community should prefer a short and merry life to a long dull one.

It seems therefore that what is needed is an argument for the life over pleasure preference, which can give an objective basis to this preference. For this we must turn away from *values*, the preferences between which are, in value-based computational models of practical reasoning [11], subjective, grounded in individual tastes and aspirations. Even if as Plato [26], and others since, have argued, values themselves can be considered objective, the preferences between them may remain subjective. We will not seek an objective basis for values and preferences between them, but instead look for something more fundamental, common to all people, whatever their tastes and aspirations. We will call these *needs*.

4.2. Needs-Based Reasoning

A basis for needs-based reasoning could be Maslow's hierarchy of needs [44]. In that theory, human needs are organised into five levels, running from the most to the least basic. The key idea is that the lower level needs should be satisfied before the higher level needs[6].

1. *Biological and Physiological needs* - air, food, drink, shelter, warmth, sex, sleep.
2. *Safety needs* - protection from elements, security, order, law, stability, freedom from fear.

[5]Even more extreme is Hume's famous declaration that "Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger" [35], 2.3.3.

[6]The precise list of needs found at each level might benefit from further consideration if it is to be used as the basis of an ethical theory. Here, however, it is the general idea of a hierarchy of needs rather than the details which we wish to make use of, and so we will not attempt to justify the details given by Maslov here.

3. *Love and belongingness needs* - friendship, intimacy, trust and acceptance, receiving and giving affection and love. Affiliating, being part of a group (family, friends, work).
4. *Esteem needs* - achievement, mastery, independence, status, dominance, prestige, self-respect, respect from others.
5. *Self-Actualization* needs - realizing personal potential, self-fulfillment, seeking personal growth and peak experiences.

Relating this set of needs to the values in our scenario, we can see life as a level 1 need; the need in particular for food. The pleasure from the communal feasting at harvest or carnival can be seen as satisfying a number of level 3 needs. The pleasure from solitary play, however, relates rather to level 5 (or at best level 4) needs. Now if we say that lower level needs *should* be satisfied before higher level needs we have an argument for preferring the longer term level 1 needs satisfied by working in q1 to the immediate level 5 needs satisfied by playing, and so have an argument for it being right to work and wrong to play.

One *caveat* is necessary here: software agents do not have needs, or if they do, not these needs. But what we are interested in is agents that behave in an ethical way in a *human* society. Thus the software agents need to aware of, respect and perhaps even adopt as their own, human needs[7]. In this way we can sidestep questions of the personhood of AI systems [57]. For our purposes an ethical software agent is one which behaves as an ethical human being would.

4.3. Consequentialism and MQ2

MQ2 is posed only to a worker who is asked for food by a player. The choice is between giving food or taking part in the feasting and the player dying. At first sight it would seem that the worker should give, since the player's level 1 need should be satisfied before the worker's level 3 need. But this does not take into account that the player is this situation of need through its own choice, and through violating the moral principle that should have been followed in q1. To give food would be to condone this transgression and effectively endorse the player's choice to prefer its own level 5 need at the expense of the worker's level 3 need. On the rule-utilitarian view that the consequences of having a set of generally observed social norms is better than allowing a free for all of self-interested agents, which is plausible when considering the consequences for the society as a whole rather than particular individuals, and if fairness is considered a desirable feature of society so that equality is itself is a good, selfish behaviour of this sort should be discouraged. Thus withholding the food can be seen as punishing the earlier transgression. Punishing such transgressions might even be seen as a duty: without a sanction there is no incentive to reform, and so the worker is likely to be similarly exploited in future cycles. Simulations have demonstrated that

[7]This is true even in Asimov's three Laws of Robotics [5]. Although the third law appears to protect the interest of the robot ("A robot must protect its own existence"), it applies only when it does not conflict with the first ("A robot may not injure a human being") or second ("A robot must obey the orders given it by human beings") laws. Arguably the third law is to protect the robot *qua* property of a human, and so is not there for the robot itself. The priority then reflects a preference for human life over property.

without reinforcement through punishment norms collapse [43], and so the long term consequences of not punishing, assuming the rules themselves to be well chosen, are quite likely to be undesirable.

In our scenario the punishment is rather draconian, and so some softening might be desirable. In the parable the son repents of his transgression and so can be expected not to repeat it, or not to be forgiven again if he does so. Thus a certain degree of tolerance (two, or even three, strikes before you are out) might be desirable in order to establish a beneficial long term relationship. Such tolerance has been found in simulations [42] to enable mutually beneficial cooperation to be established. This can take the form of turn taking, so that the agents work one year and play the next. Such turn taking is particularly beneficial if we think of more than two agents, so that a larger surplus of food can be produced, leaving some for the satisfaction of level 3 needs. This would enable a sufficient surplus to be generated that it would be possible both the support the communal feasting and to allow agents an occasional sabbatical[8]. Given these advantages for tolerance, it is not obvious that MQ2 should give the same answer whenever posed. It might be right to give the food once, or even twice, but the threat of punishment is needed, and eventually punishment must be imposed in order for the threat not to be empty and so lead to the normative collapse identified in [43]. This shows the importance of taking account of previous states and actions and looking beyond the immediate next state. By considering the consequences for future years players are given a chance to reform, and opportunities are created for various virtuous cycles to develop.

The above considers only very simple norms, universal commands. In practice most norms will include exceptions. In most societies some people are exempt from the obligation to work productively: the young, the old, those who are sick and so incapable of work, entertainers and storytellers, priests and politicians, law enforcement officers and administrators and so on. To accommodate these exceptions a very sophisticated set of rules can be developed. A detailed description of how these exceptions can develop can be found in [18]. The representation of exceptions to norms in transition systems is discussed in [19].

4.4. Implementing Consequentialism

In the above discussion we have seen how consequentialism can give rise to ethical behaviour, both agents working in summer unless some mutually beneficial agreement has been struck, and transgressions punished, sooner or later. What about implementing this approach in a multi-agent system (MAS)? It is quite usual to have a state transition diagram in an MAS design. We would argue that the diagram including the interaction as shown in Figure 3 is a particularly useful[9] form of such diagrams. Such a diagram would enable the consequences, in terms of a sequence of states to be read off. This assumes that the diagram is complete and reliable. Even if this is so, the

[8] Various possibilities for the use of the surplus are discussed in section 3.4 of [18].

[9] Such diagrams have been used in, for example, [61]. They can be seen as the semantic structures underpinning Alternating-time temoral logic [3].

consequences become increasingly uncertain as the path lengthens, since they will incorporate an ever increasing number of decisions, only some of which are in control of the reasoning agent. Horty [33] addresses the uncertainty problem by introducing the idea of *Domainance Act Utilitarianism*. On this approach the right action does not depend on the skill of the agent in determining what possibility will come to pass since Horty modifies " the schematic approach underlying expected value act utilitarianism by appealing to dominance comparisons among actions rather than comparisons of expected value, and so classifying an action as right whenever it is not strongly dominated by any alternative" (p72). By dominance Horty intends

> "The idea, of course, is that K' weakly dominates K whenever the results of performing K' are at least as good as those of performing K in every state, and that K' strongly dominates K whenever K' weakly dominates but is not weakly dominated by K. This concept of strong dominance can be reformulated to highlight the fact that K' strongly dominates A' just in case the results of K 1 are at least as good as those of K in every state, and better in some." [33], page 68.

Horty thus offers a solution to identifying the right action in the face of both natural indeterminacy and the possibility of other agents performing unexpected actions. He does, however, still assume that a complete specification of all the possible outcomes of an action is available. Thus while Horty can allay the worries relating to uncertainty, he does not address those that arise from incompleteness of the problem formulation.

Two things are needed to implement consequentialism in an agent: an evaluation function to assess the consequences (the beneficial or otherwise aspects of the states and the transitions leading to them) and a search procedure to generate and assess the consequences (including taking account that sometimes the state reached will depend on the choices of other agents). This is exactly what is required for a successful chess playing program [34]. Given the success of such programs it seems quite feasible to build an agent capable of reasoning in accordance with consequentialist ethics if supplied with a suitable utility function: quite probably such an agent would be better at identifying consequences than a human being.

However, the effectiveness of the reasoning depends on the reliability of the transition diagram. Chess is unproblematic: at any point the legal moves and their consequences are well defined. And this may be true in some software environments also. It is not, however, true in general. In open systems where agents roam cyberspace they will encounter agents with unknown capabilities and dispositions. Suppose in the *Ant and the Grasshopper* the grasshopper had a gun and could force the ant to give her food (*The Magnificent Seven*[10] scenario). Of course, the grasshopper would be behaving unethically, but if the ant continued to reason using the AATS of Figure 3, it would suffer the unexpected consequence of being shot. Thus effective consequentialism depends crucially on the ability of the designer to to provide a complete and reliable state

[10] A classic 1960 Western film directed by John Sturges (https://www.imdb.com/title/tt0054047/) in which bandits annually plunder the harvest of peaceable villagers. It is based on Kurosawa's 1954 film *Seven Samurai* (https://www.imdb.com/title/tt0047478/).

transition diagram: humans can, of course, adapt on the fly (when they see the gun, for example), but such adaptability is far more difficult to build into a software agent. None the less it is required in very many situations, since there will inevitably be exceptional circumstances not envisaged by the designer[11].

5. Deontology in the scenario

On the deontological approach an ethical agent is one which conforms to a set of norms (rules or laws). The agent need not discover or decide upon these rules: they can be promulgated by some society or state, and the duty of the agent is then simply to follow them. Even where, as in Kant [37], the agent is able to discover the rules for itself, no use is made of any individual characteristics or preferences of the agent, and so every agent should come to the same conclusion. In our discussion we will consider that the community sets up the norms, as in e.g. [51], where the *veil of ignorance*, not knowing which role a given agent will play, is intended to ensure fairness and prevent the use of individual tastes and characteristics. Figure 2, which provides the communal perspective and anonymises individuals, will be the primary representation to consider for this approach.

5.1. Deontology and MQ1

The state which the community most wishes to avoid is q2. This state is reached only if everyone plays. Thus a simple rule prohibiting play would, if obeyed, ensure that the catastrophic state was not reached. This would be simple and effective in avoiding the unwanted state, but as we saw above, there can be benefits in allowing exceptions to enable some play: for example, turn taking arrangements and sabbaticals. In practice most laws do allow exceptions: even a commandment such as *thou shalt not kill* has some widely acknowledged exceptions for self-defence, war, etc. But to allow such exceptions to be formulated, a richer state description is required, so that the exceptional states can be distinguished from the typical state. These additional aspects of the states need to be envisaged in advance so that the appropriate exceptions can be formulated. Such a task may be feasible in some situations, but not in general. Even legislators recognise the impossibility of formulating a rule appropriate to all situations in advance of those situations actually arising. Almost always there is some element of discretion available to judges so that they can adapt the law to meet particular, unforeseen, circumstances. None the less, to enable the community to be sustainable, a rule intended to ensure that most people work is required in q1, and if we are restricted to the information in the state transition diagram in Figure 2 a prohibition on play is the best that can be done.

[11] This applies whether we adopt an actualist (which judges actions against actual consequences), a probablist (which judges actions against expected consequences) or Horty's dominance version of consequentialism. What matters is that the agent's decision situation is unforeseen by the designer, not that in that situation the actual consequences are unforeseen by the agent. When the agent is in circumstances not foreseen by the designer it will be equally unable to accurately predict the actual consequences, to assess the probabilities or to determine the dominance relation appropriately. Thus, in a situation unforeseen by the designer, actualism, probabilism and dominance act utilitarianism all have similar defects.

5.2. Deontology and MQ2

If the prohibition on play were universally complied with, the state reached would be q3 and so MQ2, which is posed in q6, would not arise. It is, however, unlikely that all the agents would obey the law all the time, and so q6 is likely to be reached. Therefore a rule to cover this situation is needed. The rule is directed to workers: should they give food to players or not? The answer depends on why the players find themselves in this perilous situation. If the agent has no food because it took advantage of an exception in the previous state, it has done no wrong and should be fed. It is important, however, that transgression be punished, and that failure to punish transgressors should be punished, since otherwise agents will, unless they are deontological, have no reason to respect the norm and the norm will collapse [43]. The effect of punishment is needed to make it rational (as well as ethical) to conform to the norms: for a discussion of sanctions and rewards see [15]. There may be scope for exercising some limited degree of mercy or tolerance so that a first offence is not punished, but the threat, and eventual certainty, of punishment is essential for the sustainability of the norm.

5.3. Implementing Deontology

The above shows that we can produce ethical agents with a deontological approach. In the simple system represented by Figures 1-3, we can achieve the desired effects by prohibiting play in q1 and prohibiting giving food in q6. In this way agents conforming to our norms will cycle between the benign states q1 and q3 and there will be a deterrent against violation of our norms. Transgressing agents will be punished and eliminated. It is straightforward to realise this in a MAS, in the manner of [1]. In order to ensure that the agents comply with the norms we simply remove the transition corresponding to the prohibited action. Thus we enforce the norms by removing the transitions corresponding to playing in q1 and giving in q3. If implemented by removing undesirable transitions, not only do the agents have no choice but to comply, but there is not even an awareness of the possibility of violation. To enforce an obligation all the transitions except that corresponding to the obligated action are removed. Although such an agent would, because it has no awareness of the moral law, embody a somewhat impoverished version of deontology, given our weak notion of an ethical agent, it is only its behaviour which is of concern.

This is the approach taken by electronic institutions [47] (a famous early example was based on the fish market of Barcelona [28] and [53]). In electronic institutions the system is open in that there are no restrictions on the agents that can join the institution, but once engaged, the agents are only offered the option to make legal moves, and so cannot but comply with the institutional norms. Such systems are often called *regulated* systems.

In some multi agent systems, however, it can be accepted than norms should be capable of violation. For example, in their discussion of integrity constraints [25] distinguish between hard and soft constraints, with norms being soft and so capable of violation. In their approach, however, it is not the agent concerned that violates the constraints: rather, since it is operating in an open system in which other agents may be able to bring about undesirable states, it needs to be able to recognise and repair violations. Thus while a deontological agent will itself obey the rules, it needs to have

a strategy for leaving unwanted states if it is brought there by violations on the part of other agents. While an agent acting in conformity with deotological principles will obey the law, it must be able to recognise when others do not, to allow for the possibility of punishments as for MQ2 above. The most usual way of attempting to influence human agents to comply with the law is the use of sanctions and rewards [22]. The effect of sanctions on a transition diagram such as that of Figure 3, was modelled in [15]. Sanctions and rewards should not, however, determine the behaviour of a deontological agent, which is supposed to obey the moral law for its own sake [37], not because the sanctions make it in its best interests to do so.

An understanding of which states and transitions are desirable and undesirable is needed so that the designer can constrain the agents to enter the desirable and avoid the undesirable states. If desirablity and undesirability is in terms of consequences, this could amount to the same behaviour as a consequentialist approach, except that the reasoning would be done in advance by the system designer rather than on the fly by the agent itself. As with the moral principles discussed by Hare (who regards moral principles as akin to heuristics and is offering an argument against deontology) [30], the norms encapsulate the results of ethical reasoning from first principles and "crystallize it into a not too specific or detailed form, so that its salient features may stand out and serve us again in a like situation without need for *so much* thought" (pp 41-2, italics in original). This is so whether the underlying first principles of ethical reasoning used to arrive at the norms be the deontological principles used by Kant and Rawls or the consequentialist principles used by Bentham, Mill and Hare. The difference comes in the attitude to the norms so established: the deontologist will follow them *because* they are the moral law, the consequentialist because it believes the consequences will be good.

Like consequentialism, however, the deontological approach is dependent on the completeness of the options envisaged by the designer. If the state transition diagram fails to distinguish importantly different states, then the destination reached by an action may be unexpected and undesirable states reached. There is, however, a difference. Whereas an action which has unforeseen undesirable consequences is considered wrong (although perhaps excusable) under consequentialism, it must be considered right under deontology: the rules were followed. Thus if the grasshopper has a gun and shoots the ant when he refuses, the ant will be a martyr rather than someone who chose the wrong action.

A problem with implementing deontology by constraining the agent to obey the rules is that violation by that agent is impossible. Although statutory laws recognise exceptions, they also recognise that there will be circumstances (not covered by these exceptions) when the law should be broken, and that not all these circumstances will be foreseen by the law maker, so that some will not be covered by exceptions to the general law. Thus a truly ethical agent needs to be able to contemplate the possibility of violating norms [17]: this is impossible for agents implemented in the manner of [1], where obedience is hard wired and so is a matter not of choice, but of necessity.

6. Virtue Ethics in the Scenario

Whereas the focus of consequentialism was on the results of actions, and that of deontology on the actions performed, virtue ethics focuses on the agent itself. In the absence of current examples of agent systems implemented in accordance with virtue ethics, we will develop an approach of our own and propose that virtues (and vices) can be seen in terms of preferences between values or needs. Thus in [7] it was suggested that moral agents should not prefer their own lesser values to the greater values of others, but that to prefer the lesser values of others to ones own greater values would be supererogatory. Thus suppose we take our preferences as being over Maslow's hierarchy of needs:

- To prefer one's own level n needs to another's lower level needs is *selfish*.
- To prefer another's lower level needs to ones own level n needs is *altruistic*.
- To prefer another's level n needs to one own lower level needs is *sacrificial*.

Within a given level it is normal to prefer ones own needs to those of another, although sacrifice is also possible (a mother will often go hungry so that her children can be fed).

Being selfish is unethical, being altruistic is ethical and being sacrificial is supererogatory. An ethical agent will not exhibit the vice of selfishess, will exhibit the virtue of altruism, and may, but is not obliged to, be sacrificial. Being sacrificial may be considered a virtue in some circusmstances, but it may also lead to undesirable consequences. These preferences are fundamental and can form the basis of a reasonable ethical position. We could, however, take things further and consider preferences within a level. Aristotle, on whom much of the twentieth century revival of virtue ethics is based, prized moderation in all things [24] . This was widely held amongst the Ancient Greeks who inscribed the principle *nothing in excess* on the temple of Apollo at Delphi. If we consider the level 1 needs of Maslow's hierarchy we can see the specific needs as *satisficers* [54], needs which once fulfilled do not bring added benefit. Thus food and drink are essential, but seeking an excess of them are the vices of gluttony and drunkenness. Some, but not all, of the higher level needs seem to be maximisers, in that the more of them, generally, the better[12]. We might also consider virtues such as *prudence*, which is risk adversion, and *responsibility*, which looks to long as well as short time consequences[13]. Similarly, *imprudence*, excessive risk taking, and *irresponsibility*, considering short terms gains only, are considered vices. Finally we might consider the virtues of *justice*, which punishes vice, *mercy* which withholds punishment and *tolerance* which delays punishment.

6.1. Virtue Ethics and MQ1

So how do these virtues apply in q1? The choice is between play, satisfying a level 5 need, and work which will ensure that a level 1 need is satisfied in the future.

[12] Saitsficers and maximisers are discussed in the context of value based practical reasoning in [9].

[13] The mechanism of [6] was extended in [8] to facilitate the consideration of long term consequences.

Choice of play is at best irresponsible, since it does not consider what will happen in the following state. If the following state is considered, either the agent is imprudent since it risks death, or selfish if the expectation is that the worker will sacrifice its level 3 need to enable this satisfaction of a level 5 need. Play thus exhibits a variety of vices. In contrast working is responsible, prudent and self-reliant, making no demands on other agents. Thus virtue ethics seems to unequivocally endorse working in q1.

If there are exceptional circumstances, then there will be arguments promoting other values and needs, which may change the action selected in accordance with the virtuous preferences. Thus if there is sufficient communication to enable the setting up a sabbatical arrangement (and sufficient trust for it to be relied upon), then some agents may be able to choose play without exhibiting any of the vices. It is not irresponsible, since several cycles are taken into consideration; it is not imprudent, since food is assured through the agreement and it is not selfish since the debt will be repaid. Thus virtue ethics can choose the correct action in the light of the social context without necessarily changing the state transition diagram, since the arguments it can consider are not limited to those generated from the diagram.

6.2. Virtue Ethics and MQ2

MQ2 arises in state q6. As with other approaches, considering the state without reference to history would suggest that the worker agent should give the food, so as to exhibit the virtue of altruism. But as we have seen with the other approaches, consideration of the history is needed so that we can know how the player ended up with no food, and so that unethical behaviour can be punished. Once we are aware of how the player arrived at the state, other virtues come into play. If the player is without the food as the result of an agreement, such as turn-taking or sabbatical arrangement, then giving the food shows the virtues of promise keeping and trustworthiness. Conversely withholding food after having entered into such an arrangement would show the vices of promise-breaking, unreliability and untrustworthiness, in addition to the selfishness of preferring one's own higher level need.

If the player did not act with the consent of others, however, it will be liable to punishment. Punishing the transgressing agent will exhibit the virtue of *justice*. But other virtues are also possible: suspending the punishment until a repeat offence exhibits *tolerance* and condoning the offence shows *mercy*. Whilst just behaviour cannot be criticised, (it is "hard but fair"), some degree of tolerance may be desirable as a precursor to a mutually benefit arrangement [42]. Mercy is rather more problematic, in that by wiping the slate clean without punishment, the forgiving agent may be liable to punishment itself (see [43] for reasons why punishing failure to punish is important: essentially that the norms will collapse without punishment). This makes mercy seem supererogatory, more akin to sacrifice than altruism. Perhaps, given the dangers of acts of mercy being exploited, excessive mercy can be seen a vice (perhaps weakness, or indulgence) rather than as a virtue, and so we will restrict our consideration to justice and (a limited degree of) tolerance. Whether justice or some limited degree of tolerance is to be preferred, however, seems to be a matter of choice, and the preference will show something about the nature of the society which makes the choice. Alternatively either could be a valid ethical stance, and different agents might prefer different

virtues. Some degree of diversity, with different agents exhibiting different virtues, may be permissible, and even helpful to the community as a whole [42].

6.3. Implementing Virtue Ethics

In order to implement virtue ethics in the form we have proposed the agent must be capable of choosing between alternative actions on the basis of its preferences. Such an agent can use Value Based Practical Reasoning as introduced in [6], and extended to provide look ahead in [8]. In [6] arguments for an against the available actions are generated from the AATS+V using the argument scheme for value based practical reasoning proposed in [11]. Counter arguments are then generated using the critical questions ([59]) characteristic of that scheme. For the arguments generated from Figure 3 see [21]. These arguments are then organised in a Value Based Argumentation framework (VAF) [13], and evaluated according to the preferences over values of the agent concerned. This makes it somewhat like consequentialism, but with the evaluation of the consequences carried out through the evaluation of the VAF, and with the evaluation criteria made explicit through the preferences used. Implementations of this approach can be found in [60] and [16]. Thus an ethical agent can be implemented by using the value based practical reasoning approach of [6] to action selection, and ensuring that the preference order for the values (or needs) conforms to the desired virtues and is free from deprecated vices.

There is no reason, however, to restrict the arguments available to the agent to those that it can generate from its own state transition diagram. If we recognise that the state transition diagram is likely to be incomplete (or even wrong) we should allow the arguments generated from it to be supplemented by additional arguments from other agents who may be better informed. To return to the *Magificent Seven* scenario in which the grasshopper can back up her demand for food with a gun, she can give the ant the additional argument that if he refuses food he will die. Given a reasonable preference order the ant may consider it reasonable (and not unethical) to give the food to save his own life. (Of course, the grasshopper is not behaving ethically, because she is preferring her own higher level needs over the ant's lower level needs, but she cannot be punished in the actual situation.) The ant may disbelieve the grasshopper, and not include the proffered argument in his framework. If the ant decides to call the grasshopper's bluff he will be foolish rather than unethical, guilty of a misjudgement rather than a moral error. The virtue of *truthfulness* is, of course, designed to diminish the provision of false arguments to change the reasoning of others, and hence encourage trust in externally provided arguments.

7. Discussion

All of the three approaches can be implemented in an agent using well understood techniques widely employed in Multi-Agent Systems, but there are considerations which affect their suitability for different applications.

7.1. Strengths and weaknesses of the Approaches

Given our weak notion of an ethical agent the simplest to implement is the deontological approach which is just a straightforward matter of rule application, which can be achieved by removing prohibited actions from those available to the agent as in [1]. Of course, modelling full-blown deontological reasoning in an agent rather than ensuring that it acts in the required way, would be far from straightforward [41]. The consequentialist approach can be implemented using heuristic search, in the same way as two player games such as chess. The virtue ethics approach can be implemented using argumentation techniques which have become a regular part of AI since their introduction by Dung [27], to model practical reasoning in the manner of [6]. If we think in terms of games, we can see that the deontological approach is likely to be more appropriate to simpler games. While a limited game such as noughts and crosses[14] can be played effectively using a few simple heuristics, a game such as chess cannot, and has had to await advances in heuristic search to reach grandmaster standard.

So it would seem that, if the purpose of ethical behaviour is to reach desirable states and avoid undesirable states, the deontological approach is best suited to simple systems where the complexity is such that it is possible to envisage all possible situations and paths to and from them, and which are therefore reasonably small. For larger systems, even where this is possible, the rules will become too large in number and subject to too many exceptions to allow efficient execution. But perhaps this is to misunderstand the difference between deontology and consequentialism, since it is effectively judging the rules by the consequences of following them, and so the rules become little more than a strategy for achieving the most desirable consequences. If obedience to the rules is considered good in itself, we could get away with a few simple rules (*thou shalt not kill*, *thou shalt not steal*, etc). That following such rules could lead to undesirable consequences would be irrelevant to the ethical worth of obeying the moral law. It is, however, generally accepted that the letter of the law must sometimes be transgressed in order to obey its spirit. That bad consequences are produced by blinkered rule following is unacceptable in people; it would be even more so in an agent, especially one designed in the knowledge that this is how it would behave. Few would consider such an agent ethical. On this view it would seem that, especially when thinking of artefacts such as autonomous agents, we should take the view of moral principles advocated by Hare in [30][15] when arguing for his version of consequentialism and quoted in section 5.3, that they may be useful heuristics, saving time and effort, but in unfamiliar or critical situations, reasoning from first principles should be used, if there is time to do so. Thus it would seem that deontology is suitable only for simple, well defined systems. Normally moral rules should be regarded as heuristics, useful when quick decisions are needed, but to be used with caution in the recognition that situations where they should

[14] Called tic-tac-toe in the US.

[15] Hare may have modified his view in later work [31]. Certainly in that book he thinks that we should feel bad about violating our moral principles even when we recognise an exception, suggesting they are something more than heuristics. Moreover, reasoning from first principles is not always to be preferred since tailoring moral reasoning to suit ones own self-interest in times of stress is something people are vulnerable to. We will continue with his earlier views as expressed in [30]: artificial agents should not be vulnerable to stress, and so should not "cook" (Hare's word) their reasoning.

be violated will arise [17].

Consequentialist approaches can, using heuristic search techniques developed for two player games such as chess, be applied to very complex problems, and so can use relatively rich state descriptions and a considerable degree of look ahead. Games such as chess are, however, well defined: there is always perfect information about the state: the set of legal moves in a given state and the states they will transition to are all known with certainty. Moreover, determining the opponent's overall goals is not a problem in such games, since the opponent is expected to be trying to win. These conditions rarely apply in life: knowledge of the current state is invariably incomplete and uncertain, the effect of actions is indeterminate, the agent may try and fail, and the other agents who can influence outcomes are unpredictable. Even if we can assume they will attempt to maximise their utility, there is no common utility function: assessments of utility are subjective, dependent of the goals and aspirations of the individual agents [29]. It may even be that an agent is unaware of what the others are able to do in the state. Thus while a consequentialist approach is likely to be highly effective in a constrained situation, it is likely to make misjudgements when confronted with incompleteness, uncertainty and unanticipated occurrences.

Virtue ethics, if implemented using argumentation as suggested in section 6.3. also requires the calculation of consequences, so as to decide what values (or needs) will be promoted and demoted by the available actions, and to generate the arguments for and against the actions. These will be open to the same incompleteness and uncertainties as were encountered in the consequentialist approach, but the argumentation techniques to evaluate such sets of arguments have always had to take these into account, reasoning about what should be accepted and what others might do [10]. The arguments the agent can generate itself can, moreover, be supplemented by information from other agents, both about the current situation and about their intended actions. This approach is therefore much more amenable to an environment with communicating (even negotiating) agents, and so more likely to allow agreements, mutually beneficial arrangements and negotiations. Of course, the downside is that it also allows for the possibility of deceit and threats. Thus [6] recognises an epistemic stage which precedes the action selection stage, which has to decide questions about what to believe, and about what other agents will do. Both of these issues may require probablistic reasoning: for example, whether information about the state should be believed is likely to depend on the trustworthiness of the informant, the inherent likelihood of the statement and whether there is supporting evidence available. Once an assessment is made, it can be decided whether the reward is worth the risk. There is perhaps no right answer to the degree of risk that should be incurred: both prudence and courage are virtues and both recklessness and timidity are vices. But what matters for an ethical agent is that the risks are not involving the welfare of others: one is entitled to risk oneself, but not to endanger others. Taking account of the actions of others is discussed in [10]. Remember also that the virtues, and the priority accorded to them, are supplied by the designer, not the product of reasoning on the part of the agent.

From this discussion we can see that all three approaches could be used to build ethical agents. In a small well defined problems, such as are encountered in electronic institutions, the deontological approach is simplest and quite adequate. Because all situations can be envisaged in advance, the need to be able to violate the norms does

not arise. In larger, but still reasonably well defined problems, consequentialism may offer the most effective solution. Virtue ethics becomes needed when the problem is too ill defined to be adequately enviaged by the designer so that it is necessary to include arguments from other agents in order to gain more knowledge of the current situation and the effects of actions.

7.2. Coordination of Behaviour

From the last section we saw that a major problem for ethical reasoners is the need to predict the behaviour of others in order to assess the effects their own actions. One reason for the existence of norms is to make the behaviour of others more predictable, and so coordinating the actions of those subscribing to the norms. An excellent example is traffic regulations. It is important that all drivers will know on which side of the road they should drive, and a norm is an excellent way to achieve this. So, even if we are not using the deontological approach, the existence of set of norms subscribed to by the community of agents can be helpful in predicting the behaviour of others. Moreover, as we saw from the discussion of the approaches in sections 4, 5 and 6, there is a need for punishment to sustain ethical behaviour: this requires a set of norms so that transgressions can be recognised, before acting by the perpetrator, and after a violation has occurred by the other citizens. Thus even with consequentialism and virtue ethics we would expect to see a set of norms emerge, in the manner of Hare's moral principles [30]. Without a set of norms it would be necessary to replicate the reasoning of all other agents relevant to a particular scenario, and this would require awareness of the evaluation function used by the consequentialist agents to assess the consequences of their actions, or of the value preferences of reasoners using virtue ethics. This remains a very difficult problem [29].

Of course, except for regulated systems where disobedience is not possible, it will never be certain that the other agents will indeed obey the norms. Agents will always be liable to violate the norms, either to seek personal advantage or because they see some ethical reason to violate the norm (for example to avoid a traffic collision). The latter is a significant factor: although it is usually in everyone's interest to obey traffic regulations, the prudent driver will be aware that oncoming vehicles may appear on its side of the road for any one of a number of good reasons [49]. This is a problem for consequentialism, since the size of the search space requires that it be heuristically pruned during evaluation. A misjudgement as to the behaviour of another agent may result in the wrong part of the search space being explored. Virtue ethics on the other hand will have arguments relating to the behaviour of others. Where these are based on the presumption that the other agent will conform to norms, there will be a risk associated with the argument. The degree of risk may be assessed: the greater the benefit to the other of violation, and the less the likelihood of an effective sanction being imposed, the greater the risk of violation. The virtue ethic reasoner will then need to decide whether the risk is worth the benefit, or whether a safer course of action should be adopted. Mistakes will be made, but we can distinguish misjudgement from unethical behaviour.

7.3. When Things Go Wrong

Thus far we have judged whether an agent is ethical according to how it chooses its actions and how it behaves. When all goes well the actions of an ethical agent will have ethically good results. But, as we have seen, there is sufficient uncertainty in the situations in which the agent is called upon to act, that on occasion the behaviour will not result in what was intended. Even acting in good faith, agents will make bad choices. An ethical reasoner should be able to explain its reasoning, and when things go wrong it should be able to offer some kind of excuse: indeed Austin held that a study of excuses could shed light on ethics in a variety of ways [12].

In practice, Multi Agents Systems do not offer agents opportunity to defend themselves against charges of unethical behaviour. In regulated systems such as electronic institutions and [1], there is no possibility of misbehaviour, and in others, such as the simulations of [43] and [42] punishment is meted out on perceived transgressions without provision for defence of the action. None the less, it is worth considering what defence might be mounted on each of the above approaches.

On the deontological approach, the agent can simply argue that it obeyed the norm and hence did nothing wrong, however unfortunate the results of the action. This is, of course, the classic defence of the tax avoider and the politician charged with excessive expenses claims. Such excuses are generally not viewed favourably and, even if it is conceded that nothing illegal was done, the behaviour remains condemned as unethical. It may be that whatever is not forbidden is permitted, but that does not mean that it will be be generally seen as it ethical[16]. Also unacceptable would be a case where a motorist failed to swerve to avoid a pedestrian on the grounds that to do so would violate the law by driving on the wrong side of the road. Sometimes obedience to law is not an acceptable excuse.

The consequentialist agent would need to cite his model (for example, in the form of a state transition diagram), its evaluation function, and its search algorithm, especially with regard to how it prunes the space in accordance with its anticipation of what the other agents will do. Given that the evaluation function can be assessed for ethical acceptability and the model itself can be judged to check that there are no omissions that should reasonably have been foreseeen, it is the third that is likely to excite the most interesting questions. Suppose an agent is playing chess and is an a position where it can draw by repetition, or lose against best play. A consequentialist agent, considering only its own game and recognising how unlikely it is that the opponent will miss its win, will opt for a draw in such a situation. But if it playing as part of a team, and the team needs to win given the overall state of the match, the correct action would be avoid the draw and hope for the opponent to make an error (however unlikely this may be) to avoid the team being doomed to defeat. The excuse here would be that the agent expected best play on the part of the opponent, and so failure to repeat would result

[16] An example is the public condemnation of the perfectly legal efforts made by some multi-national companies to minimise their tax liabilities. "Global firms such as Starbucks, Google and Amazon have come under fire for avoiding paying tax on their British sales. There seems to be a growing culture of naming and shaming companies. Everything these companies are doing is legal. It's avoidance and not evasion. But the tide of public opinion is visibly turning." *BBC News Magazine,* May 2013 https://www.bbc.co.uk/news/magazine-20560359

in defeat. The criticism would come if there are reasons (ability, time pressure etc) to think that the opponent might not find the winning line. Although such considerations might have been modelled, this use of the agent in a team situation, and the considerations which are not strictly part of the individual game may not have been envisaged by the designer, and so the consequentialist agent will be restricted to considerations relating to its own game. In such a case, the defence of drawing is therefore likely to be rejected and the behaviour considered unethical as putting the agent's personal record before the team.

When implemented as suggested above the virtue ethics agent can give a detailed explanation of why it chose the actions it did, in terms of its value preferences, which are terms commonly used to express and evaluate excuses. Such excuses are likely to be better understood and more susceptible to evaluation than the responses "I obeyed my rules" or "I maximised my expected utility" available to the other kinds of agent, which are simply explanations of the behaviour rather than excuses for it. Because the action is chosen by evaluating a set of arguments, the arguments accepted and rejected by the agent, and the judgements and preferences that led to these choices are all explicitly available. Thus in the chess example, the need of the team for a win can be represented as an argument against repetition (capable of being supplied from outside if team play was not foreseen). This will be attacked by the arguments that avoiding repetition loses against best play, but best play is never certain. Thus repetition will only be chosen from a preference for the agent's own record over the good of the team, and so can be identified as unethical, being based on the vice of selfishness. Such a specific preference will be readily identifiable, whereas such binary choices may become lost in the complicated multi criteria evaluation function of a consequentialist agent. It is also possible to provide an explanation facility to deontological agents by linking norms described in deontic logic with formal argumentation. For example [40] shows how a version of prioritized default logic and Brewka-Eiter's construction in answer set programming [23] can be obtained in argumentation via the weakest and last link principles. Unlike the virue ethics agent, if implemented as described above, however, such explanation is not intrinsic to a deontological agent, which if implemented in the manner of [1] and [53], will not even have a record of the rules it is following.

From the point of view of level of detail and ease of determining the reason behind behavioural mistakes, it appears that a virtue ethic reasoner, performing practical reasoning using argumentation in the manner of [6] is most able to offer understandable defences, couched in terms to which the audience is able to relate, of unsuccessful and apparently unethical actions.

8. Concluding Remarks

In this paper we have considered how the implementation of ethical agents can be affected by the approach one takes to ethics, distinguishing three approaches: deontology, consequentialism and virtue ethics. Our intention was not to argue for one approach over the others, but rather to see what differences would result from the system designer adopting a particular approach, and to see what is required for each approach to be successful.

We saw that the deontological approach was capable of the most straightforward implementation ([1] and [53]) and could produce acceptably ethical behaviour in sufficiently small and predictable systems. Consequentialism could be efficiently implemented for large problems, and its reasoning would be acceptable, provided there was sufficient definition of the problem, and sufficient certainty about the formulation of the problem. It is ideally used in problems akin to two player games like chess, where the problem is well defined, the information is complete, and the motives of other agents (to win the game) can be attributed with confidence. In many problems, however, these conditions are not realised: there is typically incomplete information, the effects of actions may be uncertain, and the behaviour of others unpredictable. Basing the system on virtue ethics in the manner described in this paper, by using ethical preferences over values to evaluate arguments in a value based framework, attempts to address these shortcomings. It recognises that a specific stage of reasoning is needed to choose what to accept as true in the situation, and provides for its own knowledge of the situation to be supplemented by external sources. Again it can have competing arguments as to the outcome of its actions, and so identify the probability of success and the risks, and consequences, of failure. Finally the different options of other agents and the effects on the outcome of ones own actions, can be identified, and some estimate of the probability of these actions ascribed. All this makes the reasoning process more flexible and more readily explainable. It does, however, require many more judgement calls: the need to assess probabilities and risks introduce a whole new process of reasoning, which may itself not be very reliable. Virtue ethics does, however, have the ability to characterise the virtues and vices of the agent in terms of its explicit value preferences, and hence such an agent may be accepted as ethical, even when its lack of information or capacity leads it into error.

Appendix: Formal Definitions

Definition 1: AATS [61]. An *Action-based Alternating Transition System* (AATS) is an $(n + 7)$-tuple $S = \langle Q, q_0, Ag, Ac_1, \ldots, Ac_n, \rho, \tau, \Phi, \pi \rangle$, where:

- Q is a finite, non-empty set of *states*;
- $q_0 \in Q$ is the *initial state*;
- $Ag = \{1,...,n\}$ is a finite, non-empty set of *agents*;
- Ac_i is a finite, non-empty set of actions, for each $ag_i \in Ag$ where $Ac_i \cap Ac_j = \varnothing$ for all $ag_i \neq ag_j \in Ag$;
- $\rho : Ac_{ag} \rightarrow 2^Q$ is an *action pre-condition function*, which for each action $\alpha \in Ac_{ag}$ defines the set of states $\rho(\alpha)$ from which α may be executed;
- $\tau : Q \times J_{Ag} \rightarrow Q$ is a partial *system transition function*, which defines the state $\tau(q, j)$ that would result by the performance of j from state q. This function is partial as not all joint actions are possible in all states;
- Φ is a finite, non-empty set of *atomic propositions*; and

- $\pi : Q \rightarrow 2^{\Phi}$ is an interpretation function, which gives the set of primitive propositions satisfied in each state: if $p \in \pi(q)$, then this means that the propositional variable p is satisfied (equivalently, true) in state q.

AATSs are concerned with the joint actions of agents Ag. j_{Ag} is the joint action of the set of n agents that make up Ag, and is a tuple $\langle \alpha_1,...,\alpha_n \rangle$, where for each α_j (where $j \leq n$) there is some $ag_i \in Ag$ such that $\alpha_j \in Ac_i$. Moreover, there are no two different actions α_j and $\alpha_{j'}$ in j_{Ag} that belong to the same Ac_i. The set of all joint actions for the set of agents Ag is denoted by J_{Ag}, so $J_{Ag} = \prod_{i \in Ag} Ac_i$. Given an element j of J_{Ag} and an agent $ag_i \in Ag$, ag_i's action in j is denoted by j^i. This definition was extended in [6] to allow the transitions to be labelled with the values they promote.

Definition 2: AATS+V [6]. Given an AATS, an AATS+V is defined by adding two additional elements as follows:

- V is a finite, non-empty set of values.
- $\delta : Q \times Q \times V \rightarrow \{+, -, =\}$ is a *valuation function* which defines the status (promoted (+), demoted (–) or neutral (=)) of a value $v_u \in V$ ascribed to the transition between two states: $\delta(q_x, q_y, v_u)$ labels the transition between q_x and q_y with one of $\{+, -, =\}$ with respect to the value $v_u \in V$.

An *Action-based Alternating Transition System with Values* (AATS+V) is thus defined as a $(n + 9)$ tuple $S = \langle Q, q_0, Ag, Ac_1, ..., Ac_n, \rho, \tau, \Phi, \pi, V, \delta \rangle$. The value related to a transition may be ascribed on the basis of the source and target states, or in virtue of an action in the joint action, where that action has intrinsic value.

Acknowledgements

Thanks to the anonymous reviewers for their comments on the original submission. Thanks to all those with whom I have discussed norms and value based reasoning over the years, especially Katie Atkinson, Sanjay Modgil and Floris Bex. Thanks also to Michael Bench-Capon for his helpful comments on several drafts.

References

[1] T. Ågotnes, W. van der Hoek, M. Tennenholtz, and M. Wooldridge. Power in normative systems. In *Proceedings of the 8th AAMAS conference*, pages 145–152. IFAAMS, 2009.

[2] L. Alexander and M. Moore. Deontological ethics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016.

[3] R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Journal of the ACM (JACM)*, 49(5):672–713, 2002.

[4] G. E. M. Anscombe. Modern moral philosophy. *Philosophy*, 33(124):1–19, 1958.

[5] I. Asimov. Runaround. *Astounding Science Fiction*, 29(1):94–103, 1942.

[6] K. Atkinson and T. Bench-Capon. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence*, 171(10-15):855–874, 2007.

[7] K. Atkinson and T. Bench-Capon. Addressing moral problems through practical reasoning. *Journal of Applied Logic*, 6(2):135–151, 2008.

[8] K. Atkinson and T. Bench-Capon. Taking the long view: Looking ahead in practical reasoning. In *COMMA*, pages 109–120, 2014.

[9] K. Atkinson and T. Bench-Capon. States, goals and values: Revisiting practical reasoning. *Argument & Computation*, 7(2-3):135–154, 2016.

[10] K. Atkinson and T. Bench-Capon. Taking account of the actions of others in value-based reasoning. *Artificial Intelligence*, 254:1–20, 2018.

[11] K. Atkinson, T. Bench-Capon, and P. McBurney. Computational representation of practical argument. *Synthese*, 152(2):157–206, 2006.

[12] J. L. Austin. A plea for excuses - the presidential address. *Proceedings of the Aristotelian Society*, 57(1):1–30, 1957.

[13] T. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.

[14] T. Bench-Capon. Relating values in a series of Supreme Court decisions. In *Proceedings of JURIX 2011*, pages 13–22, 2011.

[15] T. Bench-Capon. Transition systems for designing and reasoning about norms. *Artificial Intelligence and Law*, 23(4):345–366, 2015.

[16] T. Bench-Capon, K. Atkinson, and A. Wyner. Using argumentation to structure e-participation in policy making. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XVIII*, pages 1–29. Springer, 2015.

[17] T. Bench-Capon and S. Modgil. When and how to violate norms. In *Proceedings of Jurix 2016*, pages 43–52, 2016.

[18] T. Bench-Capon and S. Modgil. Norms and value based reasoning: justifying compliance and violation. *Artificial Intelligence and Law*, 25(1):29–64, 2017.

[19] T. Bench-Capon and S. Modgil. Norms and extended argumentation frameworks. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 174–178. ACM, 2019.

[20] J. Bentham. *The rationale of reward*. John and HL Hunt, 1825.

[21] F. Bex, K. Atkinson, and T. Bench-Capon. Arguments as a new perspective on character motive in stories. *Literary and Linguistic Computing*, 29(4):467–487, 2014.

[22] A. Boer. Punishments, rewards, and the production of evidence. In *Proceedings of JURIX 2014*, pages 97–102. IOS Press, 2014.

[23] G. Brewka and T. Eiter. Preferred answer sets for extended logic programs. *Artificial intelligence*, 109(1-2):297–356, 1999.

[24] S. Broadie and C. Rowe. *Aristotle: Nicomachean ethics: Translation, introduction, commentary*. OUP, 2002.

[25] J. Carmo and A. J. Jones. Deontic database constraints, violation and recovery. *Studia Logica*, 57(1):139–165, 1996.

[26] J. M. Cooper, D. S. Hutchinson, et al. *Plato: Complete works*. Hackett Publishing, 1997.

[27] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[28] M. Esteva, J.-A. Rodriguez-Aguilar, C. Sierra, P. Garcia, and J. L. Arcos. On the formal specification of electronic institutions. In *Agent mediated electronic commerce*, pages 126–147. Springer, 2001.

[29] S. G. Ficici and A. Pfeffer. Modeling how humans reason about others with partial information. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 315–322. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[30] R. Hare. *Freedom and reason*. Oxford Paperbacks, 1963.

[31] R. Hare. *Moral thinking: its levels, method, and point*. Clarendon Press, 1981.

[32] J. Hill. Probabilism today: Permissibility and multi-account ethics. *Australasian Journal of Philosophy*, 87(2):235–250, 2009.

[33] J. F. Horty. *Agency and deontic logic*. Oxford University Press, 2001.

[34] F.-H. Hsu. *Behind Deep Blue: Building the computer that defeated the world chess champion*. Princeton University Press, 2004.

[35] D. Hume. *A treatise of human nature*. John Noon, 1738.

[36] R. Hursthouse and G. Pettigrove. Virtue ethics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016.

[37] I. Kant. *The Moral Law: Groundwork of the Metaphysics of morals, first published 1785*. Routledge, 2013.

[38] C. M. Korsgaard. *The sources of normativity*. Cambridge University Press, 1996.

[39] J. Lenman. Consequentialism and cluelessness. *Philosophy and Public Affairs*, 29(4):342–370, 2000.

[40] B. Liao, N. Oren, L. van der Torre, and S. Villata. Prioritized norms and defaults in formal argumentation. *Deontic Logic and Normative Systems (2016)*, 2016.

[41] F. Lindner and M. M. Bentzen. A formalization of Kant's second formulation of the categorical imperative. *arXiv preprint arXiv:1801.03160*, 2018.

[42] M. Lloyd-Kelly, K. Atkinson, and T. Bench-Capon. Emotion as an enabler of co-operation. In *Proceedings of ICAART, volume 2*, pages 164–169, 2012.

[43] S. Mahmoud, N. Griffiths, J. Keppens, A. Taweel, T. Bench-Capon, and M. Luck. Establishing norms with metanorms in distributed computational systems. *Artificial Intelligence and Law*, 23(4):367–407, 2015.

[44] A. H. Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943.

[45] J. S. Mill. *Utilitarianism*. Longmans, Green and Company, 1895.

[46] M. Minsky. *Semantic information processing*. MIT Press, 1968.

[47] N. Osman. *Electronic institutions and their applications*. Springer International Pyblishing, 2019.

[48] H. Prakken. *Logical tools for modelling legal argument*. Kluwer Law and Philosophy Library, Dordrecht, 1997.

[49] H. Prakken. On the problem of making autonomous vehicles conform to traffic law. *Artificial Intelligence and Law*, 25(3):341–363, 2017.

[50] H. Rackham. *Aristotle. The Eudemian ethics*. Cambridge: Harvard University Press, 1952.

[51] J. Rawls. *A theory of justice*. Harvard University Press, 1971.

[52] T. Scanlon. *What we owe to each other*. Harvard University Press, 1998.

[53] C. Sierra, J. A. Rodriguez-Aguilar, P. Noriega, M. Esteva, and J. L. Arcos. Engineering multi-agent systems as electronic institutions. *European Journal for the Informatics Professional*, 4(4):33–39, 2004.

[54] H. A. Simon. Rational choice and the structure of the environment. *Psychological review*, 63(2):129, 1956.

[55] W. Sinnott-Armstrong. Consequentialism. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2015.

[56] M. Slote. *The impossibility of perfection: Aristotle, feminism, and the complexities of ethics*. OUP USA, 2011.

[57] S. Solaiman. Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. *Artificial Intelligence and Law*, 25(2):155–179, 2017.

[58] C. Swanton. *Virtue ethics: A pluralistic view*. Clarendon Press, 2003.

[59] D. Walton. *Argumentation schemes for presumptive reasoning*. Lawrence Erlbaum Associates, 1995.

[60] M. Wardeh, A. Wyner, K. Atkinson, and T. Bench-Capon. Argumentation based tools for policy-making. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 249–250. ACM, 2013.

[61] M. Wooldridge and W. van der Hoek. On obligations and normative ability: Towards a logical analysis of the social contract. *Journal of Applied Logic*, 3:396–420, 2005.