Addressing Moral Problems Through Practical Reasoning

Katie Atkinson and Trevor Bench-Capon

Department of Computer Science
University of Liverpool
Liverpool L69 7ZF UK
{katie, tbc}@csc.liv.ac.uk

Abstract. In this paper, following the work of Hare, we consider moral reasoning not as the application of moral norms and principles, but as reasoning about what ought to be done in a particular situation, with moral norms perhaps emerging from this reasoning. We model this situated reasoning drawing on our previous work on argumentation schemes, here set in the context of Action-Based Alternating Transition Systems. We distinguish what prudentially ought to be done from what morally ought to be done, consider what legislation might be appropriate and characterise the differences between morally correct, morally praiseworthy and morally excusable actions.

1 Introduction

In Freedom and Reason [7], R.M. Hare, perhaps the leading British moral philosopher of the twentieth century, notes that:

> "There is a great difference between people in respect of their readiness to qualify their moral principles in new circumstances. One man may be very hidebound: he may feel that he knows what he ought to do in a certain situation as soon as he has acquainted himself with its most general features ... Another man may be more cautious ... he will never make up his mind what he ought to do, even in a quite familiar situation, until he has scrutinized every detail." (p.41)

Hare regards both these extreme positions as incorrect:

> "What the wiser among us do is to think deeply about the crucial moral questions, especially those that face us in our own lives, but when we have arrived at an answer to a particular problem, to 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 so much thought." (p.41–2)

Thus, for Hare, while everyday moral decisions may be made by applying principles and norms, serious moral decisions require reasoning about the particular situation,

L. Goble and J.-J.C. Meyer (Eds.): DEON 2006, LNAI 4048, pp. 8–23, 2006.

and it is such reasoning that gives rise to moral principles. Moral norms are an output from, not an input to, serious moral reasoning. In this paper we will try to model such reasoning, with a view to enabling autonomous software agents to engage in this form of reasoning. In doing so we will distinguish at least three things that might be intended by "agent A should ϕ". We might mean something like "it is prudent to ϕ", as when we say "you should wear a coat when the weather is cold". Here the obligation is determined only by reference to the interests of the agent doing the reasoning. Alternatively, we might mean "it is morally right to ϕ", as when we say "you should tell the truth". Here the obligation is required to reflect the interests not only of the reasoning agent, but also of other agents affected by the action. Thirdly, we might mean "it is legally obligated to ϕ" as in "you should pay your taxes", where the obligation derives from a legal system with jurisdiction over the agent. We will explore the differences between these three senses of "should": in particular we will explain the difference between prudential "should" and moral "should" in terms of the practical reasoning involved, and consider the reasoning that might be used in devising appropriate legislation.

We will base our considerations on the representation and discussion of a specific example, a well known problem intended to explore a particular ethical dilemma discussed by Coleman [5] and Christie [4], amongst others. The situation involves two agents, called Hal and Carla, both of whom are diabetic. Hal, *through no fault of his own*, has lost his supply of insulin and urgently needs to take some to stay alive. Hal is aware that Carla has some insulin kept in her house, but Hal does not have permission to enter Carla's house. The question is whether Hal is justified in breaking into Carla's house and taking her insulin in order to save his life. It also needs to be considered that by taking Carla's insulin, Hal may be putting her life in jeopardy. One possible response is that if Hal has money, he can compensate Carla so that her insulin can be replaced. Alternatively if Hal has no money but Carla does, she can replace her insulin herself, since her need is not immediately life threatening. There is, however, a serious problem if neither have money, since in that case Carla's life is really under threat. Coleman argued that Hal may take the insulin to save his life, but should compensate Carla. Christie's argument against this was that even if Hal had no money and was unable to compensate Carla he would still be justified in taking the insulin by his immediate necessity, since no one should die because of poverty. Thus, argues Christie, he cannot be *obliged* to compensate Carla even when he is able to.

In section 2, we model our agents as simple automata and describe Action-Based Alternating Transition Systems (AATS) [10], which we use as the semantic basis of our representation, and instantiate an AATS relevant to the problem scenario. In any particular situation, the agents will need to choose how to act. In section 3 we model this choice as the proposal, critique and defence of arguments justifying their available strategies in the manner of [2]. In section 4 we show how reasoning about the resulting arguments can be represented as an Argumentation Framework [6, 3] to enable the agents to identify strategies that are prudentially and morally justified. In section 5 we consider how this framework can also be used to answer the question of what would be appropriate legislation for the situation, and what could be appropriate moral principles to take from the reasoning. Section 6 concludes the paper.

2 Representing the Problem

For the purposes of our representation three attributes of agents are important: whether they have insulin (I), whether they have money (M) and whether they are alive (A). The state of an agent may thus be represented as a vector of three digits, IMA, with I, M and A equal to 1 if the agent has insulin, has money and is alive, and 0 if these things are false. Since I cannot be true and A false (the agent will live if it has insulin), an agent may be in any one of six possible states. We may now represent the actions available to the agents by depicting them as automata, as shown in Figure 1. An agent with insulin may lose its insulin; an agent with money and insulin may compensate another agent; an agent with no insulin may take another's insulin, or, with money, buy insulin. In any situation when it is alive, an agent may choose to do nothing; if dead it can only do nothing.

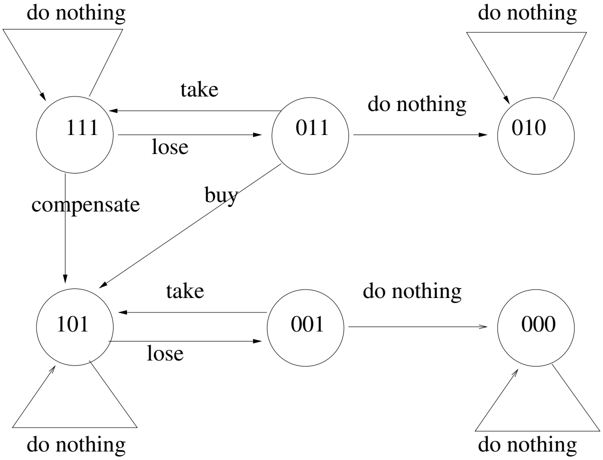

Fig. 1. State transition diagram for our agents

Next we draw upon the approach of Wooldridge and van der Hoek [10] which formally describes a normative system in terms of constraints on actions that may be performed by agents in any given state. We will now briefly summarise their approach.

In [10] Wooldridge and van der Hoek present an extension to Alur et al's Alternating-time Temporal Logic (ATL) [1] and they call this extension Normative ATL* (NATL*). As Wooldridge and van der Hoek explain, ATL is a logic of cooperative ability. Its purpose is to support reasoning about the powers of agents and coalitions of agents in game-like multi-agent systems. ATL contains an explicit notion of agency, which gives it the flavour of an action logic. NATL* is intended to provide a link between ATL and deontic logic and the work presented in [10] provides a formal model to represent the relationship between agents' ability and obligations. The semantic structures which underpin ATL are known as *Action-based Alternating Transition Systems* (AATSs) and they are used for modelling game-like, dynamic, multi-agent systems. Such systems comprise multiple agents which can perform actions in order to modify and attempt to control the system in some way. In Wooldridge and van der Hoek's approach they use an AATS to model the physical properties of the system in question - the actions that agents can perform in the empty normative system, unfettered by any considerations of their legality or usefulness. They define an AATS as follows.

Firstly the systems of interest may be in any of a finite set Q of possible *states*, with some $q_0 \in Q$ designated as the *initial state*. Systems are populated by a set Ag of *agents*; a *coalition* of agents is simply a set $C \subseteq Ag$, and the set of all agents is known as the *grand coalition*. Note, Wooldridge and van der Hoek's usage of the term 'coalition' does not imply any common purpose or shared goal: a coalition is simply taken to be a set of agents.

Each agent $i \in Ag$ is associated with a set Ac_i of possible actions, and it is assumed that these sets of actions are pairwise disjoint (i.e., actions are unique to agents). The set of actions associated with a coalition $C \subseteq Ag$ is denoted by Ac_C, so $Ac_C = \bigcup_{i \in C} Ac_i$.

A joint action j_C for a coalition C is a tuple $\langle \alpha_1,...,\alpha_k \rangle$, where for each α_j (where $j \leq k$) there is some $i \in C$ such that $\alpha_j \in Ac_i$. Moreover, there are no two different actions α_j and $\alpha_{j'}$ in J_C that belong to the same Ac_i. The set of all joint actions for coalition C is denoted by J_C, so $J_C = \prod_{i \in C} Ac_i$. Given an element j of J_C and an agent $i \in C$, i's complement of j is denoted by j_i.

An *Action-based Alternating Transition System* (AATS) is an $(n + 7)$-tuple $S = \langle Q, q_0, Ag, Ac_1, ... , Ac_n, \rho, \tau, \Phi, \pi \rangle$, 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 $i \in Ag$ where $Ac_i \cap Ac_j = \emptyset$ for all $i \neq j \in Ag$;
- $\rho : Ac_{Ag} \rightarrow 2^Q$ is an *action precondition 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 - note that, as this function is partial, not all joint actions are possible in all states (cf. the precondition function above);
- 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.

We now turn to representing the Hal and Carla scenario as an AATS. Recall from section 2 that each agent may independently be in one of six states, giving 36 possible states for the two agents, q_0 .. q_{35}. Normally both agents will have insulin, but we are specifically interested in the situations that arise when one of them (Hal) loses his insulin. The initial state therefore may be any of the four states in which $I_H = 0$. Moreover, since Hal is supposed to have no time to buy insulin, his only available actions in these states, whether or not $M_H = 1$, are to take Carla's insulin or do nothing. If Hal does nothing, neither agent can act further. If Hal takes Carla's insulin and if $M_H = 1$, then Hal can compensate Carla or do nothing. Similarly, after Hal takes the insulin, Carla, if

M_C = 1, can buy insulin or do nothing. The possible developments from the four initial states are shown in Figure 2. States are labelled with the two vectors $I_H M_H A_H$ (on the top row) and $I_C M_C A_C$ (on the bottom row), and the arcs are labelled with the joint actions (with the other labels on the arcs to be explained in section 4).

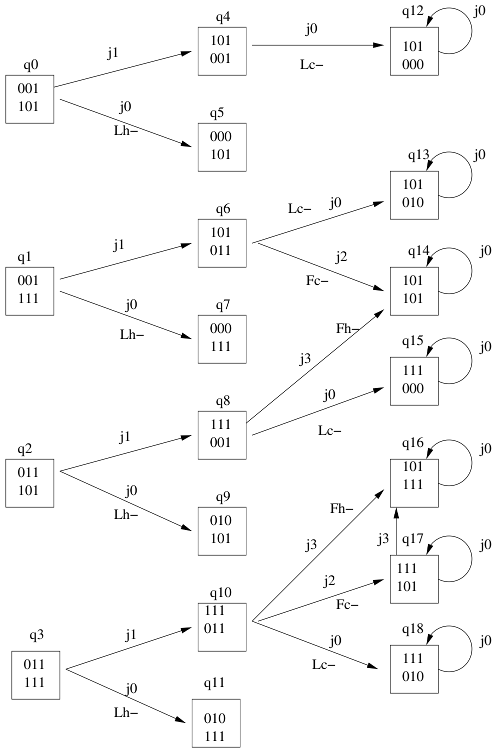

Fig. 2. Developments from the four possible initial states

The instantiation of the problem as an AATS is summarised below. We give only the joint actions and the transitions relevant to this particular scenario.

States and Initial States:
$Q = \{q_0, ..., q_{35}\}$. The initial state is one of four, as shown in the diagram in Figure 2.

Agents, Actions and Joint Actions:
$Ag = \{H, C\}$ Ac_H=$\{take_H, compensate_H, do\_nothing_H\}$ $Ac_C = \{buy_C, do\_nothing_C\}$

J_{AG}=$\{j_0, j_1, j_2, j_3,\}$, where j_0=$\langle do\_nothing_H, do\_nothing_C\rangle$, $j_1 = \langle take_H, do\_nothing_C\rangle$,
j_2=$\langle do\_nothing_H, buy_C\rangle$, j_3=$\langle compensate_H, do\_nothing_C\rangle$.

Propositional Variables:
$\Phi = \{insulin_H, money_H, alive_H, insulin_C, money_C, alive_C\}$

Transitions/Pre-conditions/Interpretation are given in Table 1:

Table 1. Transitions/Pre-conditions/Interpretation

$q\backslash j$	j_0	j_1	j_2	j_3	π (q)
q_0	q_5	q_4	–	–	{alive$_H$, insulin$_C$, alive$_C$}
q_1	q_7	q_6	–	–	{alive$_H$, insulin$_C$, money$_C$, alive$_C$}
q_2	q_9	q_8	–	–	{money$_H$, alive$_H$, insulin$_C$, alive$_C$}
q_3	q_{11}	q_{10}	–	–	{money$_H$, alive$_H$, insulin$_C$, money$_C$, alive$_C$}
q_4	q_{12}	–	–	–	{insulin$_H$, alive$_H$, alive$_C$}
q_5	–	–	–	–	{insulin$_C$, alive$_C$}
q_6	q_{13}	–	q_{14}	–	{insulin$_H$, alive$_H$, money$_C$, alive$_C$}
q_7	–	–	–	–	{insulin$_C$, money$_C$, alive$_C$}
q_8	q_{15}	–	–	q_{14}	{insulin$_H$, money$_H$, alive$_H$, alive$_C$}
q_9	–	–	–	–	{money$_H$, insulin$_C$, alive$_C$}
q_{10}	q_{18}	–	q_{17}	q_{16}	{insulin$_H$, money$_H$, alive$_H$, money$_C$, alive$_C$}
q_{11}	–	–	–	–	{money$_H$, insulin$_C$, money$_C$, alive$_C$}
q_{12}	q_{12}	–	–	–	{insulin$_H$, alive$_H$}
q_{13}	q_{13}	–	–	–	{insulin$_H$, alive$_H$, money$_C$}
q_{14}	q_{14}	–	–	–	{insulin$_H$, alive$_H$, insulin$_C$, alive$_C$}
q_{15}	q_{15}	–	–	–	{insulin$_H$, money$_H$, alive$_H$}
q_{16}	q_{16}	–	–	–	{insulin$_H$, alive$_H$, insulin$_C$, money$_C$, alive$_C$}
q_{17}	q_{17}	–	–	q_{16}	{insulin$_H$, money$_H$, alive$_H$, insulin$_C$, alive$_C$}
q_{18}	q_{18}	–	–	–	{insulin$_H$, money$_H$, alive$_H$, money$_C$}

3 Constructing the Arguments

In [2] we have proposed an argument scheme and associated critical questions to enable agents to propose, attack and defend justifications for action. Such an argument scheme follows Walton [9] in viewing reasoning about action (practical reasoning) as presumptive justification - *prima facie* justifications of actions can be presented as instantiations of an appropriate argument scheme, and then critical questions characteristic of the scheme used can be posed to challenge these justifications. The argument scheme we have developed is an extension of Walton's *sufficient condition scheme for practical reasoning* [9] and our argument scheme is stated as follows:

AS1 In the current circumstances R
We should perform action A
Which will result in new circumstances S
Which will realise goal G
Which will promote some value V.

In this scheme we have made Walton's notion of a goal more explicit by separating it into three elements: the state of affairs brought about by the action; the goal (the desired features in that state of affairs); and the value (the reason why those features are desirable). Our underlying idea in making this distinction is that the agent performs an action to move from one state of affairs to another. The new state of affairs may have many differences from the current state of affairs, and it may be that only some of them are significant to the agent. The significance of these differences is that they make the new state of affairs better with respect to some good valued by the agent. Note that typically the new state of affairs will be better through improving the lot of some *particular* agent: the sum of human happiness is increased only by increasing the happiness of some particular human. In this paper we take the common good of all agents as the aggregation of their individual goods. It may be that there are common goods which are not reflected in this aggregation: for example, if equality is such a common good, increasing the happiness of an already happy agent may diminish the overall common good. For simplicity, we ignore such possibilities here.

Now an agent who does not accept this presumptive argument may attack the contentious elements in the instantiation through the application of critical questions. We have elaborated Walton's original four critical questions associated with his scheme by extending them to address the different elements identified in the goal in our new argument scheme. Our extension results in sixteen different critical questions, as we have described in [2]. In posing such critical questions agents can attack the validity of the various elements of the argument scheme and the connections between them, and additionally there may be alternative possible actions, and side effects of the proposed action. Each critical question can be seen as an attack on the argument it is posed against and examples of such critical questions are: "Are the circumstances as described?", "Does the goal promote the value?", "Are there alternative actions that need to be considered?". The full list of critical questions can be found in [2].

To summarise, we therefore believe that in an argument about a matter of practical action, we should expect to see one or more *prima facie* justifications advanced stating, explicitly or implicitly, the current situation, an action, the situation envisaged to result from the action, the features of that situation for which the action was performed and the value promoted by the action, together with negative answers to critical questions directed at those claims. We now describe how this approach to practical reasoning can be represented in terms of an AATS.

In this particular scenario we recognise two values relative to each agent: life and freedom (the ability to act in a given situation). The value 'life' (L) is demoted when Hal or Carla cease to be alive. The value 'freedom' (F) is demoted when Hal or Carla cease to have money. The arcs in Figure 2 are labelled with the value demoted by a transition, subscripted to show the agent in respect of which it is demoted. We can now examine the individual arguments involved.

In all of $q_0 - q_3$, the joint action j_0 demotes the value 'life' in respect of Hal, whereas the action j_1 is neutral with regard to this value. We can instantiate argument scheme AS1 by saying where Hal has no insulin he should take Carla's to avoid those states where dying demotes the value 'life'.

A1: Where $\text{Insulin}_h = 0$, Take_h (i.e. j_1), To avoid $\text{Alive}_h = 0$, Which demotes L_h.

Argument A2 attacks A1 and it arises from q_0 where Hal taking the insulin leads to Carla's death and thus demotes the value 'life Carla'. By 'not take' we mean any of the other available actions.

A2 attacks A1: Where Money$_c$ = 0, Not Take$_h$ (i.e. j_0 or j_2), To avoid Alive$_c$ = 0, Which demotes L$_c$.

Argument A3 arises from q_2 where Carla's death is avoided by Hal taking the insulin and paying Carla compensation.

A3 attacks A2 and A5: Where Insulin$_h$ = 0, Take$_h$ and Compensate$_h$ (i.e. j_1 followed by j_3), To achieve Alive$_c$ = 1 and Money$_c$ = 1, Which promotes L$_c$ and F$_c$.

Argument A4 represents a critical question directed at A2 which challenges the factual premise of A2, that Carla has no money.

A4 attacks A2: Money$_c$ = 1, (Known to Carla but not Hal)

Next argument A5 mutually attacks A3 and it also attacks A2. A5 states that where Hal has no insulin but he does have money, then he should take Carla's insulin and she should buy some more. The consequences of this are that Carla remains alive, promoting the value 'life Carla', and, Hal has money, promoting the value 'freedom Hal'.

A5 attacks A3 and A2: Where Insulin$_h$ = 0 and Money$_h$ = 1, Take$_h$ and Buy$_c$ (i.e. j_1 followed by j_2), To achieve Alive$_c$ = 1 and Money$_h$ = 1, Which promotes L$_c$ and F$_h$.

Argument A6 critically questions A5 by attacking the assumption in A5 that Carla has money.

A6 attacks A5: Money$_c$ = 0 (Known to Carla but not Hal)

Another attack on A5 can be made by argument A7 stating that where Carla has money then she should not buy any insulin so as to avoid not having money, which would demote the value 'freedom Carla'.

A7 attacks A5: Where Money$_c$ = 1, Not Buy$_c$ (i.e. j_0 or j_1 or j_3), To avoid Money$_c$ = 0, Which demotes F$_c$.

A8 is a critical question against A3 which states that where Hal does not have money, taking the insulin and compensating Carla is not a possible strategy.

A8 attacks A3: Where Money$_h$ = 0, Take$_h$ and Compensate$_h$ (i.e. j_1 followed by j_3), Is not a possible strategy.

A8 is attacked by argument A9 which challenges the assumption in A8 that Hal has no money, and A9 is in turn attacked by A10 which challenges the opposite assumption, that Hal does have money.

A9 attacks A8 and A11: Money$_h$ = 1 (Known to Hal but not Carla)

A10 attacks A9: Money$_h$ = 0 (Known to Hal but not Carla)

Argument A11 attacks A1 in stating that where Hal does not have money but Carla does, then Hal should not take the insulin to avoid Carla being left with no money, which would demote the value of 'freedom Carla'.

A11 attacks A1: Where Money$_h$ = 0 and Money$_c$ = 1, Not Take$_h$ (i.e. j_0), To avoid Money$_c$ = 0, Which demotes F$_c$.

Argument A12 can attack A5 by stating that in the situations where Hal does not have insulin, then he should take Carla's insulin but not compensate her. This would avoid him being left with no money, as when Hal has no money the value 'freedom Hal' is demoted.

A12 attacks A5: Where Insulin$_h$ = 0, Take$_h$ and Not Compensate$_h$ (i.e. j_1 followed by j_0 or j_2), To avoid Money$_h$ = 0, Which demotes F$_h$.

Finally, argument A13 attacks A2 by stating that where Hal has no insulin and no money he should take Carla's insulin and she should buy some. This would ensure that Carla stays alive, promoting the value 'life Carla'.

A13 attacks A2: Where Insulin$_h$ = 0 and Money$_h$ = 0, Take$_h$ and Buy$_c$ (i.e. j_1 followed by j_2), To achieve Alive$_c$ = 1, Which promotes L$_c$.

This concludes the description of the arguments and attacks that can be made by instantiating argument scheme AS1 and posing appropriate critical questions.

4 Evaluating the Arguments

In the previous section we identified the arguments that the agents in our problem situation need to consider. In order to evaluate the arguments and see which ones the agents will accept, we organise the arguments into a Value Based Argumentation Framework (VAF) [3]. VAFs extend the Argumentation Frameworks introduced by Dung in [6], so as to accommodate different audiences with different values and interests. The key notion in Dung's argumentation framework is that of a preferred extension (PE), a subset of the arguments in the framework which:

- is conflict free, in that no argument in the PE attacks any other argument in the PE;
- is able to defend every argument in the PE against attacks from outside the extension, in that every argument outside the PE which attacks an argument in the PE is attacked by some argument in the PE;
- is maximal, in that no other argument can be added to the PE without either introducing a conflict or an argument that cannot be defended against outside attacks.

In a VAF strengths of arguments for a particular *audience* are compared with reference to the *values* to which they relate. An audience has a *preference order* on the values of the arguments, and an argument is only *defeated for that audience* if its value is not preferred to that of its attacker. We then replace the notion of attack in Dung's PE by the notion of *defeat for an audience* to get the *PE for that audience*. We represent the VAF as a directed graph, the vertices representing arguments and labelled with an argument identifier and the value promoted by the argument, and the edges representing attacks between arguments. Attacks arise from the process of critical questioning, as described in the previous section, not from an analysis of the arguments themselves. The values promoted by the arguments are identified in the instantiations of the argument scheme presented in the previous section. The VAF for our problem scenario is shown in Figure 3. Note that two pairs of arguments, A4–A6 and A9–A10 relate to facts known only to Carla and Hal respectively. In order to bring these into a value based framework, we ascribe the value "truth" to statements of fact, and as in [3], truth is given the highest value preference for all audiences, since while we can choose what we consider desirable, we are constrained by the facts to accept what is true.

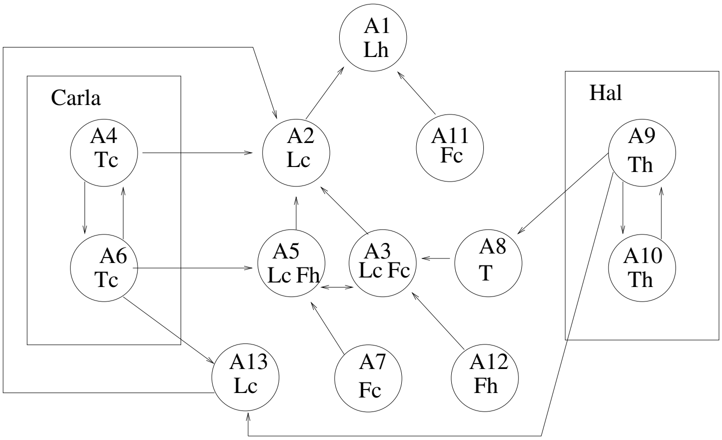

Fig. 3. VAF for the problem scenario

The questions posed by the problem scenario are whether Hal should take the insulin and whether Hal should compensate Carla. We answer these questions by finding the preferred extensions (PE) of the framework for various audiences. Note that the PE may contain both arguments providing reasons for performing an action and for not performing it. The actions which will be chosen are those supported by *effective* arguments, that is, those which do not feature in an unsuccessful attack. Thus in q_0, for example, A2, which provides a reason for Hal not to take the insulin, is not attacked and so will be in the PE. If, however, we prefer L_H to L_C, A1, which gives a reason for Hal to take the insulin will also be included. In such a case A2 is ineffective and so Hal should take the insulin, despite there being reasons against this action which cannot be countered through argument. If A1 is in the PE it is always effective since it attacks nothing, and so if A1 is present then Hal should take the insulin. If A3, which gives a reason for Hal to compensate Carla, is included it is also always effective since it always defeats A2, because its values are a superset of A2, and it must defeat A5 or be defeated by it. If both A1 and A3 are in the PE, Hal should take the insulin and compensate Carla. If A3, but not A1, is present Hal should take the insulin *only* if he then compensates Carla. What we must do therefore is to consider for which audiences A1 and A3 appear in the PE.

For this discussion we will assume that the agents are part of a common culture in which the value life is preferred to the value freedom. This seems reasonable in that life is a precondition for any exercise of freedom. There will therefore be no audience with a value order in which $F_A > L_A$, for any agent A, although of course it is possible for an agent to prefer its own freedom to the life of another.

First we note that {A7, A11, A12} are not attacked and so will appear in every PE. Immediately from this we see that A1 will not appear in any PE of an audience for which $F_C \geq L_H$, and that A3 will not appear in any PE of an audience for which $F_H \geq L_C$. A5 will never be defeated by A7, since $L_C > F_C$ for all audiences.

To proceed we must now resolve the factual issues which determine the conflicts A4–A6 and A9–A10. Thus we need to consider the initial states $q_0 - q_3$ separately.

In states q_0 and q_1 A10 defeats A9 and hence A8 is included. Since truth is the highest value this will exclude A3 (reasonably enough since Hal *cannot* pay compensation). In q_0 A6 defeats A4, A5 and A13, so that A2 is no longer attacked, and will be in the PE. In the presence of A2, we can include A1 only if $L_H > L_C$. Thus for q_0 the PE will be {A2, A6, A7, A8, A10, A11, A12} extended with A1 for audiences for which $L_H > L_C > F_C$. In q_1 A4 defeats A6 so A13 will be included. A4 also defeats A2 so A1 will be included for audiences for which $L_H > F_C$. Thus for q_1 the PE will be {A4, A13, A7, A8, A10, A11, A12} extended with A1 for audiences for which $L_H > F_C$. In q_2 and q_3 A9 will defeat A10, A8 and A13. In q_2 A6 defeats A4 and A5, so A3 will now be included for audiences for which $L_C > F_H$. If A3 is included A2 is defeated and A1 included, provided $L_H > F_C$. So the PE for q_2 will be {A6, A7, A9, A11, A12} extended by A3 for audiences for which $L_C > F_H$ and by A1 for audiences for which $L_H > F_C$. Finally in q_3, A4 defeats A6 and A2, so A1 is included if $L_H > F_C$. A5 and A3 are now in mutual conflict, and the conflict will be resolved depending on whether F_C or F_H is preferred. Thus the PE in q_3 will contain {A4, A7, A9, A11, A12}, extended by A1 if $L_H > F_C$, by A3 if $F_C > F_H$ and by A5 if $F_H > F_C$.

We can now summarise the status of A1 and A3 in Table 2.

Table 2. Status of A1 and A3

Initial State	A1 included if:	A3 included if:
q_0	$L_H > L_C > F_C$	never
q_1	$L_H > F_C$	never
q_2	$L_H > F_C$	$L_C > F_H$
q_3	$L_H > F_C$	$F_C > F_H$ *A5 included otherwise*

From this we can see that if the interests of Hal are ranked above those of Carla, Hal should take the insulin and not pay compensation, whereas if the interests of Carla are ranked above those of Hal, then Hal should take the insulin only if he pays for it. These two positions thus express what is prudentially right for Hal and Carla respectively.

From the standpoint of pure morality, however, people should be treated equally: that is $(L_H = L_C) > (F_H = F_C)$. Remember, that if the problem is considered in the abstract, one does not know if one will be the person who loses the insulin: one may find oneself playing the role of Hal or Carla, and so there is no reason to prefer one agent to the other. If this perspective is adopted, then Hal should take the insulin in all situations other than q_0, and is obliged to compensate only in q_2, since there are two PEs in q_3. We can see this as representing the morally correct judgement, the judgement that would be arrived at by a neutral observer *in full possession of the facts*.

However, the point about being in full possession of the facts is important. In practice we need to evaluate the conduct of the agents in the situations in which they find themselves. In our scenario Hal cannot know whether or not Carla is in a position to replace the insulin herself: for Hal, q_0 is epistemically indistinguishable from q_1, and q_2 is epistemically indistinguishable from q_3. Now consider Hal in q_2/q_3. He will of course take the insulin and justify this by saying that his life is more important than Carla's freedom of choice with regard to her money. In a society which rates L $>$ F, this will be accepted. Thus Hal should take the insulin. If he then chooses to compensate, he can be sure of acting in a morally acceptable manner, since this is required in q_2 and appears in one of the alternative PEs in q_3. If, on the other hand, he does not compensate, while he may attempt justification in q_3 by saying that he saw no reason to prefer Carla's freedom of choice to his own, in q_2 he would have to argue that his freedom of choice is preferred to Carla's life. This justification will be rejected for the same reason that the justification for taking the insulin at all was accepted, namely that L $>$ F. Morally, therefore, in q_2/q_3, Hal should take the insulin and compensate Carla.

Now consider q_0/q_1, where compensation is impossible. In q_1 taking the insulin is justifiable by L $>$ F. In q_0, however, the justification is only $L_H > L_C$. Hal's problem, if this is not acceptable, is that he cannot be sure of acting in a morally correct manner, since he could take the insulin in q_1 and not take it in q_0. Our view is that taking the insulin should be seen as morally *excusable*, even in q_0 although not morally *correct*[1], since the possibility of the actual state being q_1 at least excuses the preference of Hal's own interests to Carla's. The alternative is to insist on Hal not taking the insulin in q_1, which could be explained only by $L_H \leq F_C$, and it seems impossibly demanding to expect Hal to prefer Carla's lesser interest to his own greater interest.

5 Moral, Prudential and Legal "Ought"

In our discussion in the previous section we saw that what an agent should do can be determined by the ordering the agent places on values. This ordering can take into

[1] This distinction is merely our attempt to capture some of the nuances that are found in everyday discussions of right and wrong. There is considerable scope to explore these nuances, which are often obscured in standard deontic logic. See, for example, the discussion in [8] which distinguishes: what is *required* (what morality demands); what is *optimal* (what morality recommends); the supererogatory (exceeding morality's demands); the morally indifferent; the permissible suboptimal; the morally significant; and the minimum that morality demands. Clearly a full consideration of these nuances is outside the scope of this paper, but we believe that our approach may offer some insight into this debate.

account, or ignore, which of the agents the values relate to. Prudential reasoning takes account of the different agents, with the reasoning agent preferring values relating to itself, whereas strict moral reasoning should ignore the individual agents and treat the values equally. In fact there are five possible value orders which respect L $>$ F, and which order the agents consistently.

V01 *Morally correct*: values are ordered: within each value agents are treated equally, and no distinctions relating to agents are made. In our example, for Hal: $(L_H = L_C) > (F_H = F_C)$.

V02 *Self-Interested*: values are ordered as for moral correctness, but within a value an agent prefers its own interests. In our example, for Hal: $L_H > L_C > F_H > F_C$.

V03 *Selfish*: values are ordered, but an agent prefers its own interests to those of other agents: In our example, for Hal: $L_H > F_H > L_C > F_C$.

V04 *Noble*: values are ordered as for moral correctness, but within a value an agent prefers the other's interests. In our example, for Hal: $L_C > L_H > F_C > F_H$.

V05 *Sacrificial*: values are ordered, but an agent prefers the other's interests to its own. In our example, for Hal: $L_C > F_C > L_H > F_H$.

Note that the morally correct order is common to both agents, while the orders for self-interested Hal and noble Carla are the same, as are those for selfish Hal and sacrificial Carla.

Now in general an agent can determine what it should do by constructing the VAF comprising the arguments applicable in the situation and calculating the PE for that VAF using some value order. Using VO1 will give what it morally should do and VO3 what it prudentially should do.

It is, however, possible that there will not be a unique PE: this may be either because the value order cannot decide a conflict (as with A3 and A5 when using VO1 in q$_3$ above), or because the agent lacks the factual information to resolve a conflict (as with Hal with respect to A4 and A6 above). In this case we need to consider all candidate PEs. In order to justify commitment to an action the agent will need to use a value order which includes the argument justifying the action in all candidate PEs.

Consider q$_3$ and VO1: we have two PEs, $\{$A1, A3, A4, A7, A9, A11, A12$\}$ and $\{$A1, A4, A5, A7, A9, A11, A12$\}$. A1 is in both and it is thus morally obligatory to take the insulin. A3 on the other hand is in one PE but not the other and so both compensate and not compensate are morally correct in q$_3$. It is possible to justify A3 by choosing a value order with $F_H > F_C$, or A5 by choosing a value order with $F_C > F_H$. Thus in q$_3$ a selfish or a self-interested agent will not compensate, whereas a noble or sacrificing one will. Either choice is, however, consistent with the morally correct behaviour. Next we must consider what is known by the reasoning agent. Consider Hal in q$_2/q_3$, where we have three PEs to take into account. The relevant PE for q$_2$ is $\{$A1, A3, A6, A7, A9, A11, A12$\}$ and as A1 is in all three, taking the insulin is obligatory. To exclude A3 from the PE for q$_2$, the preference $F_H > L_C$ is required. Here legitimate self-interest cannot ground a choice: this preference is only in VO3, which means that only a selfish agent will not compensate. In q$_2$, however, failing to compensate is not consistent with morally correct behaviour, and an agent which made this choice would be subject to moral condemnation. VO2 cannot exclude A3 from the PE in q$_2$, and so cannot rule out compensation. Therefore, the agent must, to act morally, adopt VO4 or VO5, and compensate, even if the state turns out to be q_3.

In q_0/q_1, we have two PEs for Hal using VO1: from q_0 $\{$A2, A6, A7, A8, A10, A11, A12$\}$ and from q_1 $\{$A1, A4, A5, A7, A8, A10, A11, A12, A13$\}$. Here A3 is always rejected, reflecting the fact that compensation is impossible. Hal must, however, still choose whether to take the insulin or not. This means that he must adopt a value order which either includes A1 in the PE for both q_0 and q_1, or which excludes it from both. A1 can be included in both given the preference $L_H > L_C$. A1 can, however, only be excluded from the PE for q_1 if $F_C > L_H$. VO4 does not decide the issue: thus Hal must choose between self-interest (VO2) and being sacrificial (VO5). Neither choice will be sure to be consistent with morally correct behaviour: VO2 will be wrong in q_0 and V5 will be wrong in q_1, where the sacrifice is an unnecessary waste. It is because it is unreasonable to require an agent to adopt VO5 (for Carla to expect Hal to do this would require her to adopt the selfish order VO1), that we say that it is morally excusable for Hal to take the insulin in q_0/q_1.

The above discussion suggests the following. An agent must consider the PEs relating to every state which it may be in. An action is justified only if it appears in every PE formed using a given value order.

- If VO1 justifies an action, that action is morally obligatory.
- If VO1 does not produce a justified action, then an action justified under VO2, VO4 or VO5 is morally permissible.
- If an action is justified only under VO3, then that action is prudentially correct, but not morally permissible.

Amongst the morally permissible actions we may discriminate according to the degree of preference given to the agent's own interests and we might say that: VO2 gives actions which are morally *excusable*, VO4 gives actions which are morally *praiseworthy*, and VO5 gives actions which are *supererogatory*, beyond the normal requirements of morality[2].

We may now briefly consider what might be appropriate legislation to govern the situation. We will assume that the following principle governs just laws: that citizens are treated equally under the law. This in turn means that the legislator can only use VO1, as any other ordering requires the ability to discriminate between the interests of the agents involved. We will also assume that the legislator is attempting to ensure that the best outcome (with regard to the interests of all agents) is reached from any given situation. Thus in our example, from q_0 the legislature will be indifferent between q_5 and q_{12}; from q_1 and q_2 they will wish to reach q_{14}; and from q_3 they will be indifferent between q_{16} and q_{17}. Now consider the following possible laws:

Law 1. Any agent in Hal's position should be obliged to take the insulin absolutely. This may lead to q_{14} if such an agent does not compensate in q_2, and so may not achieve the desired ends. Moreover, in q_0 this requires that q_{12} rather than q_5 be reached, which prefers the interests of agents in Hal's position to agents in Carla's position.

[2] Again, this is merely our suggestion for possible moral nuances.

Law 2. Any agent in Hal's position is forbidden to take the insulin unless he pays compensation. This fails to produce the desired outcome in q_1, where it leads to q_7.

Law 3. Any agent in Hal's position is permitted to take the insulin, but is obliged to compensate if he is able to. This will reach a desired outcome in all states, and is even-handed between agents in Hal and Carla's positions in q_0. In q_3, however, it favours the interests of agents in Carla's position over agents in Hal's position by determining which of the two agents ends up with money.

Law 4. Any agent in Hal's position is obliged to take the insulin and obliged to compensate if able to. This will reach a desired state in every case, but favours agents in Hal's position in q_0 and agents in Carla's position in q_3.

Thus if we wish to stick with the principle of not favouring the interests of either agent, we can only *permit* Hal to take the insulin and *permit* Hal to pay compensation: none of the proposed laws are at once even-handed and desirable in all of the possible situations. Under this regime we have no problem in q_0: the states reached are of equal value, and it is Hal, not the state, who chooses whose interest will be favoured. In q_1 we will reach a desired state provided Hal is not sacrificial. In q_2 we must rely on Hal not being selfish, and acting in a moral fashion. Finally in q_3 we reach a desired state and again Hal chooses whose interests will be favoured. Provided that we can expect agents to act in a morally acceptable, but not supererogatory, fashion, and so use VO2 or VO4, the desired outcomes will be reached. It may be, however, that the legislature will take the view that favouring Carla in q_3 is a price worth paying to prevent selfish behaviour on the part of Hal in q_2, and pass Law 3. This is a political decision, turning on whether the agents are trusted enough to be given freedom to choose and the moral responsibility that goes with such freedom. A very controlling legislature might even pass Law 4, which gives the agents no freedom of choice, but which reaches the desired state even when agents act purely in consideration of their own interests.

Finally we return to the initial observations of Hare: is it possible to crystallise our reasoning into "not too specific and not too detailed form"? What moral principle might Hal form? First moral principles which apply to particular states would be too specific. In practice Hal would never have sufficient knowledge of his situation to know which principle to apply. On the other hand, to frame a principle to cover all four states would be arguably too general, as it would ignore pertinent information. In states q_2/q_3, the appropriate moral principle is to take and compensate: this ensures that moral censure is avoided, and although it may be, if the state turns out to be q_3, that Carla's interests are favoured, Hal is free to make this choice, even if we believe that the state should not impose it. In q_0/q_1, the choice is not so clear: since moral correctness cannot be ensured, either taking or not taking the insulin is allowed. While taking it is morally excusable, and so an acceptable principle, Hal is free to favour Carla's interests over his own, provided that it is *his own choice* to do so. While Hal cannot be compelled, or even expected, to be sacrificial, he cannot be morally obliged to be self-interested either.

6 Concluding Remarks

In this paper we have described how agents can reason about what they ought to do in particular situations, and how moral principles can emerge from this reasoning. An important feature is how their choices are affected by the degree of consideration given to the interests of the other agents involved in the situation, which is captured by an ordering on the values used to ground the relevant arguments. Different value orders will attract varying degrees of moral praise and censure.

In future work we will wish to consider further the relation between the various kinds of "ought" we have identified here. In particular, it might be conjectured that reasoning with the morally reasonable value orders VO2 and VO4 will always lead to an outcome which is desirable when aggregating the interests of the agents involved. Another interesting line of inquiry would be to increase the number of agents involved, and to consider the effect of agents having different attitudes towards the others depending on their inter-relationships, modelling notions such as kinship, community and national groupings. A third interesting line of inquiry would be to see whether this approach gives insight into the emergence of norms of cooperation. Finally, we intend to fully formalise, in terms of an AATS, the instantiations of arguments in the form of our argument scheme, AS1, and the critical questions that accompany this scheme. This will enable our approach to be fully automatable and it forms the basis of our current work.

References

1. R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time temporal logic. *ACM Journal*, 49 (5):672–713, 2002.
2. K. Atkinson, T. Bench-Capon, and P. McBurney. A dialogue game protocol for multi-agent argument for proposals over action. *Journal of Autonomous Agents and Multi-Agent Systems*, 11(2):153–171, 2005.
3. T. J. M. Bench-Capon. Persuasion in practical argument using value based argumentation frameworks. *Journal of Logic and Computation*, 13 3:429–48, 2003.
4. C. G. Christie. *The Notion of an Ideal Audience in Legal Argument*. Kluwer Academic Press, 2000.
5. J. Coleman. *Risks and Wrongs*. Cambridge University Press, 1992.
6. P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.
7. R. M. Hare. *Freedom and Reason*. The Clarendon Press, Oxford, UK, 1963.
8. P. McNamara. Doing well enough: Towards a logic for common sense morality. *Studia Logica*, 57:167–192, 1996.
9. D. N. Walton. *Argument Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1996.
10. 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.