Modelling State Intervention

Alison Chorley and Trevor Bench-Capon
Department of Computer Science
The University of Liverpool
Liverpool
UK

Abstract

In previous work we have looked at how agents reason in situations of conflict of interest, using an approach based on an argument scheme for practical reasoning and associated critical questions. In this paper we add the possibility of the State intervening to attempt to improve the outcome. We model the State as another agent in the scenario, with its own repertoire of actions, and its own interests represented as an ordering on values. Where arguments are directed towards different agents with their own different interests it is not possible to use standard means of determining the acceptability of arguments, since these methods evaluate arguments from a single perspective. We therefore adopt the approach of simulating the reasoning of the agents using a procedural version of the argument scheme and associated questions. We present our work through consideration of an extended example, draw some conclusions and identify directions for future work.

1. Introduction

In [ABC] we considered how agents might employ practical reasoning in order to decide how to act in the case of a particular moral dilemma. In [CBC] we addressed the same problem empirically via an implementation of a procedural version of the approach. In these papers we established that the actions chosen by the agents depended on ordering of the values subscribed to by the agents, and that certain orderings could give rise to globally undesirable outcomes. In this paper we will consider how the State (we use *State* for the ruling authority, to distinguish it from the *states* of the system) might intervene in the situation in order to ensure the best outcome. Our approach will be to model the State as a third agent, with its own repertoire of actions and its own values.

The ability of the State to manipulate the outcome of the situation was briefly discussed in [ABC], where the State was restricted to either obligating or prohibiting certain actions. There we showed that under the constraint that any such prohibition or obligation had to be even handed as between the citizens of the State, no such law could be formulated, since every plausible formulation favoured one or other the agents in certain circumstances. It was there argued that to achieve a satisfactory outcome it was necessary to rely on the agents behaving in a morally acceptable manner, that is considering the interests of the other agent to a reasonable degree, by adopting an appropriate ordering on their values. Here we will extend our notion of the State to consider an active State that can influence the situation directly instead of simply constraining the behaviour of its citizens.

In section 2 we will recapitulate the example moral dilemma, and extend it to model the State. Section 3 will describe our representation of the problem. In section 4 we will consider the practical arguments relating to the actions of the State, and suggest their resolution. In section 5 we will investigate the situation empirically so as to show the effects of State intervention. Finally, in section 5, we will give some concluding remarks.

2 The Example problem

We 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 [Col] and Christie [Chr], amongst others. The situation involves two agents, 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 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. 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 she can replace her insulin. Alternatively if Carla has money, she can replenish her insulin herself. There is, however, a serious problem if neither have money, since in that case Carla's life is really under threat.

In this paper we will extend the example by giving the State two possible actions: it can give a person insulin, and it can fine a person who takes the property of another. These two actions should suffice to ensure that even in the worst case where neither Hal nor Carla have money, both should live.

3. Representation

We represent our example as we did in [ABC] and [CBC] as an Action Based Alternating Transition System (AATS) [WvdH], a structure based on Alternating Time Temporal Logic [AHK], extended to include the notion of the values of an agent. Agents have states, represented as a vector of propositions, a repertoire of actions, which change the states of agents, and a set of values, used to assess the worth of a transition between states. The states of the AATS are the aggregate of the individual states of the agents in the system, and transitions between these states are by means of *joint actions*, that is a set of actions, one for each agent. Each of these transitions can be labelled to show whether the move to the new state promotes or demotes a value for each agent. Here we will use the instantiation of the AATS given in [CBC].

The state of the citizen agents, Hal and Carla, consist of a four tuple <I,M,A,W>, where I is 0 if the citizen has no insulin and 1 if the citizen has insulin. M indicates the financial state of the citizen, 0 for no money, 1 for exactly enough money to buy insulin, and 2 for more than enough money. A indicates the health of the citizen: 0 for dead, 1 for in immediate need of insulin, and 2 for in good health. Finally, W indicates the states of the world: 1 if shops are open and so insulin can be bought, and 0 otherwise. If the citizen has insulin, then health is good: if I = 1 then A = 2. W is the same for both agents.

The actions of the citizens are that they may take another's insulin, buy insulin, compensate another citizen by transferring a unit of money, or do nothing. The values of the citizen are life and freedom (which requires that $A > 0$, and is then increased as M increases). Since these values are promoted or demoted by reference to the health or wealth of a *particular* agent, we subscript them with the name of the agent, and subscripted values are regarded as distinct.

We next need to include the State. We assume its supplies of insulin and money are effectively unlimited, but it is important to represent expenditure. We thus introduce an additional proposition R (reserves) which is 1 if there was no change from the previous state, 0 if reserves decreased, and 2 if reserves increased. The State may give a citizen insulin: we add as a precondition that $W = 1$, since otherwise no application can be made: giving insulin will set R to 0. The state may also fine a citizen, reducing M for that citizen by 1, and setting R to 2. Like its citizens, the State values their lives and their freedom, but it is also concerned with its own financial state, which is its *own* degree of freedom, giving rise to the additional value, F_S.

In the initial state of the example for Hal, $I= 0$ and $A = 1$, for Carla $I =1$, $A =2$, and $W =0$. The example does not specify the wealth of Hal or Carla, and so there are four possible initial states, as M can be either 0 or 1 for each of them. An example state transition diagram is shown in Figure 1.

4. Constructing Arguments

For generating arguments, we use the argument scheme and associated critical questions developed in [ATK] as an extension of Walton's sufficient condition scheme [WAL] for practical reasoning. The agent will determine what it is best to do by considering the options and justifications and resolving conflicts according to the ranking it places on the justifying values. The basic scheme is

In state R
Do action A
To reach next state S
In which goal G is true
And G promotes value V

The scheme also has a negative version:

In state R
Refrain from action A
To avoid next state S
In which goal G is true
And G demotes value V

Finally there is a mixed version:

In state R
Do action A
To avoid next state S
In which goal G is true

And G demotes value V

The critical questions challenge various aspects of an instantiation of this scheme, claiming things such as the current state is different, the action will not reach the intended state, the value will not be promoted, that some other value will be demoted, etc. Formal definitions of the argument scheme and critical questions in terms of the AATS structure are given in [ABC2]. Since it this representation goals are simply a subset of propositions true in a state, there is never a dispute as to whether a given state realises a goal or not, we will, in this paper, collapse state and goal into a single element.

We now apply this approach to the situation where a citizen has applied for insulin. We begin by considering only the action of giving insulin, which allows us to consider the simpler AATS in which only the applicant and the state are represented. We begin with an argument to supply the insulin:

A1: Where $A_C = 1$, $I_C = 0$, $W = 1$
State should give insulin to citizen
To avoid $A_C = 0$
Which demotes L_C

First we must question whether the initial state is as described. We can therefore frame three attacks, any of which would suffice to defeat A1:

Attack 1: Not $I_C = 0$
Attack2: Not $A_C = 1$
Attack 3: Not $W = 1$

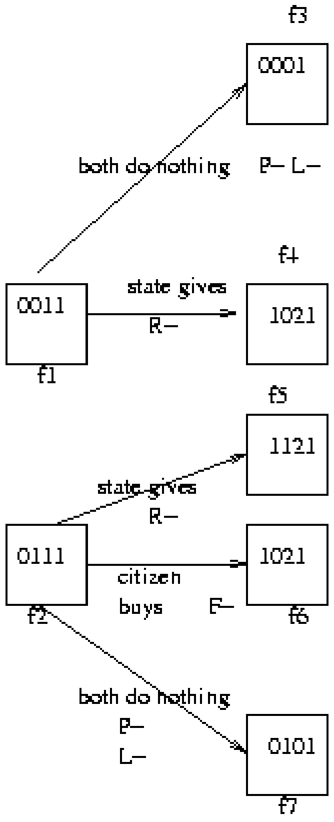

Figure 1: State transitions for problem situation

We will, however, assume that these factual aspects are satisfied: this simply imposes a restriction on the set of initial states that we need to consider. We now look at further arguments which arise is these states. Of course, giving the insulin demotes the value reserves, and so we get the counter argument

A2: Where $A_C = 1$, $I_C = 0$, $W = 1$
State should not give insulin to citizen
To avoid $R = 0$
Which demotes F_S

We may now suppose that the citizen has money and so can by insulin.

A3: Where $A_C = 1$, $I_C = 0$, $M_C > 0$, $W = 1$
Citizen should buy insulin
To avoid $R=0$
Which demotes F_S

A3 is in conflict with A1, since the citizen will remain alive anyway if she buys the insulin herself. This, of course, requires that the citizen does have money and so is open to the objection

Attack4: $M_C = 0$

Since our problem did not specify the value of M, this attack must be considered. The citizen could, however, object to A3 with A4:

A4: Where $A_C = 1$, $I_C = 0$, $M_C > 0$, $W = 1$
Citizen should not buy insulin
To avoid M_C decreased
Which demotes F_C

We can arrange these arguments into a Value based Argumentation Framework [BC], as shown in Figure 2.

Now let us consider the possibility of fines. For this we need to include the state of the other agent, and to add another variable, O (for Owes), which is set to 1 when Hal takes the insulin and back to zero if and when Hal compensates.

We can give an argument in favour of fining Hal:

A5: Where $A_C = 1$, $I_C = 0$, $W = 1$, $M_H > 0$ and $O_H = 1$
State should fine Hal
To increase R
Which promotes F_S

This argument attacks A2, since now reserves will not be reduced, even if the State does give the insulin. A5 can, of course, be attacked by the suggestion that Hal has no money.

Attack5: $M_H = 0$

Additionally there is an argument against fining

A6: Where $A_C = 1$, $I_C = 0$, $W = 1$, $M_H > 0$ and $O_H = 1$
State should not fine
To avoid decrease in M_H
Which demotes F_H

We can arrange these arguments into a Value based Argumentation Framework [BC], as shown in Figure 2. We cannot, however, apply one of the standard techniques to calculate the status of the arguments in this framework by determining the admissible sets because there are a multiplicity of agents involved. This while, for example, A^ is directed towards the State, A4 is directed towards the citizen. Since these different agents are more than likely to have different interests, their orderings on values will differ. This in turn means that there is no single audience with respect to which we can evaluate the status of the arguments.

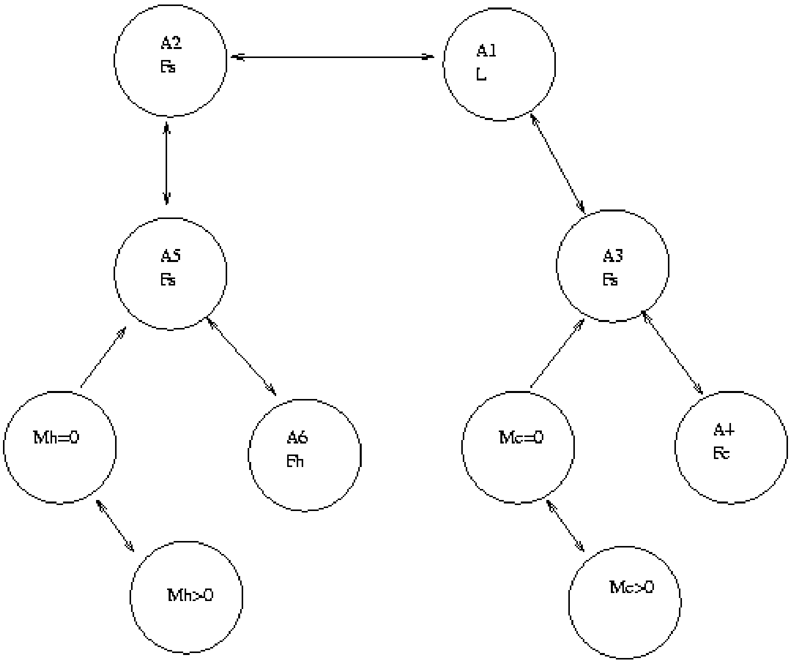

Figure 2: VAF with State actions

This means that we will have to adopt a different approach, one in which it is possible to accommodate a plurality of agents, each with their own interests and values. Our idea is to simulate the reasoning of the agents involved in the scenario, in which each will decide on its own best course of action with respect to its own preferences. This approach will be described in the next section.

5. An Empirical Approach to the Problem

We use for our empirical exploration the program reported in [CBC], extended to allow the modelling of several agents. This program relies on a procedural interpretation of the practical argument scheme and critical questions approach described above.

The argument schemes described in section 4 are instantiated giving a set of arguments for each agent with one or more arguments per transition, depending on how many values are affected by a transition. The agents now each order the arguments directed towards their actions, beginning with the argument promoting their most favoured value down to that promoting that least favoured value, through any neutral arguments to the argument demoting its least favoured value and finally to the argument demoting its most favoured value. Each argument will be considered in turn, as providing the currently best presumptive justification, until one that can be defended against the relevant critical questions is reached.

The program uses three critical questions to put to the arguments, to test whether that can be defended against possible drawbacks:

[PCQ1] Might the action lead to states that the agent will wish to avoid?
[PCQ2] Might the other agent fail to act so as to perform the desired joint action?
[PCQ3] Is the desired state in fact a local optimum, so that all subsequent states will result in a state worse than the current one?

PCQ1 relates to whether we have a stronger argument against performing the action. This argument may be from an unfortunate side effect of the target state itself, in that it demotes a value we prefer to the one it promotes. Remember, however, that the state we actually reach from performing an action may not be the one we wish to reach, since the chosen action only determines a set of joint actions. Thus the choice of the other agent may mean that performing a particular action will take us to an unfavourable state: this risk can only be avoided by refraining from the action.

The rebuttal to PCQ1 involves considering the arguments available to the other agent. On the assumption that the other agent is rational, it will be reasoning in a similar fashion (note that this may mean that we need to make some assumptions about the value preferences of that agent. Now if the other agent also has a reason to avoid the undesired state, we can discount the risk. Thus if the other agent has available an argument instructing it to avoid the undesired state, we may consider rejecting PCQ1. PCQ1, however, may be re-instated if the other agent has a counter-rebuttal: that is if the other agent has a better reason (in terms of its own value ordering) to reach the undesired state. In this case we must consider PCQ1 unanswered and reject the argument it attacks.

PCQ2 also involves the other agent. In this case the other agent may have a reason for avoiding the state we wish to reach. In this case, there is no point in acting to reach the state since we will expect the other agent to frustrate our attempt. The rebuttal to PCQ2 is that the other agent has a stronger reason to reach the state we desire. Given such an argument we may expect it to cooperate and participate in the joint action which will reach this state.

PCQ3 arises from the possibility that the state we are trying to reach may be initially promising, but ultimately lead to unfortunate consequences. Thus we may have a reason to avoid a state, even if it promotes a value, if all subsequent choices that can be made in that state will result in us being worse off than we were in the initial state. This involves looking ahead to some later state. In the case where paths do not terminate, some cut-off to keep within resource bounds must be applied. Again the rebuttal of this question involves the other agent having a compelling argument to avoid the state with this property, and no stronger argument to reach it.

We can now run this program, taking as the initial states those where Hal has lost his insulin. There are four such states to consider, according to whether Hal, Carla, both or neither has money. We will consider the case where both the citizen agents have a *selfish* value order: $L_{self} > F_{self} > L_{other} > F_{other}$. In [ABC] it was shown that problems arise in two cases: where neither have money so that there is not enough insulin to go round and so one will die, and in the case where Hal, but not Carla, has money, but Hal does not compensate her, leaving her unable to buy insulin. Thus it is the situation where we have selfish that we need the State to intervene.

For the State there are three possible value orders, assuming that it respects the citizen's preference for life over freedom.

SO 1. $L > F_S > F$
SO 2. $L > F > F_S$
SO 3. $F_S > L > F$

Now that the State is committed to intervention, it may be that it needs to prefer one citizen over the other, since any course of action would express a preference between them. We will assume that the State will prefer Carla (as the "innocent" party) in such cases. Executing the program gives the results shown in Table 2. In every case Hal will take the insulin, and will not compensate Carla.

Table 2: Actions for the different situations

Preference	$M_C=0$		$M_C >0$	
	$M_H = 0$	$M_H > 0$	$M_H = 0$	$M_H > 0$
$L > F_S > F$	give insulin	fine + give insulin	give insulin	fine + give insulin
$L > F > F_S$	give insulin	give insulin	give insulin	give insulin
$F_S > L > F$	nothing	fine	buy insulin	fine + buy

When neither Hal or Carla has money the program will give insulin to Carla and not fining Hal when the State prefers Life to its Financial Interests and by not giving insulin and not fining in the reverse situation. Thus the State is capable, where it desires to do so, of ensuring that the unfortunate situation where one of the agents dies can be avoided. Where the State is reluctant to spend its resources to save its citizen, it does nothing. When Hal has money and Carla does not the State will fine Hal and then give insulin to Carla when the first value order is used and the State prefers its Financial Interests over the agents Finances. Where it prefers all the other agent values over its Financial Interests, the State will give insulin but will not fine Hal. For the third value order when the State prefers its Financial Interests over everything, the

State only fines Hal and does not give insulin to Carla, effectively punishing Hal, but doing nothing to ameliorate the consequences of his action.

When Carla has money and Hal does not the same outcomes are obtained as for when they both have no money in that the State gives insulin and does not fine when Life is preferred over its Financial Interests and does nothing in the reverse situation. The only difference here is that Carla has money and so can buy insulin when the State does not intervene. Note, however, that even where the State prefers its own financial interests to those of its citizens, it gives the insulin rather than requiring the citizen to buy. This is because the citizen is aware that the State will not allow her to die, and so forces it to give the insulin by refusing to by it herself.

Finally, in the situation where both agents have money, for the first value order the State will fine Hal and give the insulin to Carla. For the second value order the State gives Carla insulin but not fine Hal. For the third value order the State will fine Hal, but still force Carla to buy her own insulin. In this case she will do so because she is aware that the State will not save her with these preferences.

In summary, when the State prefers its Financial Interests over all the other values it will not intervene and save Carla's life by giving her insulin and will fine Hal if he has money. Carla only survives if she has her own money with which to buy insulin. When the state prefers Life to its Financial Interests it will always give insulin to Carla even when she has money to buy it. When Hal has money he fined unless the State is willing to incur expense to maintain the financial state of its citizens.

6. Discussion

A number of points of interest arise from this work. First there is the observation that Carla can force the State to supply insulin to her, even on the first value order when it would rather not. Of course, the problem could be addressed by adding an extra precondition to giving insulin, to check whether the recipient was able to buy insulin themselves, effectively "means testing" the benefit. This is, however, desirable only on one of the value orderings, so that the State would need to be clear as to its value priorities before taking this step.

Second, there is the need to know the value order of the other agents. If Carla were to believe the State was acting according to one of the first two orderings when in fact it adopted the third ordering, she would be making a fatal error. It is, of course, a principle that the State promulgates its legislation to its citizens so that they should be clear as to how it will respond in various situations. For other citizens the ascription of selfish order ("rational" as it is termed in economics"), is probably the safest assumption.

Third we should note that the simulation does not do anything to induce Hal to behave in a better way. Hal never compensates, even when he will be fined. We would like the intervention of the State to have an impact on the behaviour of the citizens, in particular for Hal to voluntarily compensate Carla directly rather indirectly through the State giving Carla insulin and fining Hal. One way would be to make the fine punitive rather than compensatory, so that it is in Hal's financial interest to compensate rather than be fined. Another possibility would be if the citizens attached some value to obeying the law. We could therefore introduce a third value for the citizens, R (Respect for Law), demoted if $O_C = 1$. Either of these would give Hal a reason to compensate Carla rather than wait to be fined, provided that he does not believe that the state prefers his freedom to its own resources. In this case, Hal is probably correct not to compensate, since the State is effectively endorsing this choice. In the other cases, Hal would still need to give a sufficient priority to this new value to act upon it. In future work will pursue this idea further.

6. Concluding Remarks.

In this paper we have described a way of modelling the intervention of the State in a particular problem scenario arising from a moral choice. We have modelled the State as a practical reasoner, acting so as to promote its own values, just as its citizens do. In order to do this we cannot rely on the standard analysis of the status of arguments in an argumentation framework, since there are multiple agents and multiple interests and perspectives that need to be taken into account. We have therefore simulated the reasoning of the agents involved, using a procedural version of the argument scheme and critical question approach used in [ABC]. A further advantage is that this procedural approach allows for larger and more refined descriptions of the problem that were used in [ABC].

We believe that this simulation approach provides insight into, and clarification of, how the State should respond to a particular situation, and have identified the role played by the values the State wishes to promote and endorse. We will explore this topic further in future work, by looking at more sophisticated scenarios, and exploring the possibility of the State’s actions influencing the conduct of its citizens.

References

[AHK] Rajeev Alur, Thomas A. Henzinger, Orna Kupferman: Alternating-time temporal logic. J. ACM 49(5): 672-713 (2002)

[ATK[K. Atkinson (2005): *What Should We Do?: Computational Representation of Persuasive Argument in Practical Reasoning.* PhD Thesis, Department of Computer Science, University of Liverpool, Liverpool, UK.

[ABC] K. Atkinson and T. Bench-Capon (2006): Addressing moral problems through practical reasoning. In L. Goble and J-J. Ch. Meyer (editors): *Deontic Logic and Artificial Normative Systems*, Lecture Notes in Artificial Intelligence (LNAI) 4048, pp. 8-23, Springer, Berlin, Germany.

[ABC2] K. Atkinson and T. Bench-Capon (2007): Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence.* Vol. 171 (10-15), pp. 855-874.

[BC] T.J.M. Bench-Capon, (2003). Persuasion in Practical Argument Using Value Based Argumentation Frameworks. *Journal of Logic and Computation*. Volume 13 No 3 pp429-448.

[CBC] A. Chorley, T. Bench-Capon, and P. McBurney. Automating Argumentation for Deliberation in Cases of Conflict of Interest. (2006). *In Proceedings of 1st International Conference on Computational models of Argument.* IOS Press. pp279-290.

[Chr] Trevor J. M. Bench-Capon: George C. Christie, The Notion of an Ideal Audience in Legal Argument. Artif. Intell. Law 9(1): 59-71 (2001)

[Col] J. Coleman. *Risks and Wrongs*. Cambridge University Press, 1992.
[WAL] D. N. Walton. *Argument Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1996.
[WvdH] M.Wooldridge andW. van der Hoek. On obligations and normative ability: Towards a logical analysis of the social contract. *Journal of Applied Logic*, 3:396–420, 2005.