Co-ordination and Co-operation in Agent Systems: Social Laws and Argumentation

Katie Atkinson and Trevor Bench-Capon

Department of Computer Science
University of Liverpool
Liverpool L69 3BX, UK
{katie,tbc}@csc.liv.ac.uk

Abstract. The social laws paradigm represents an important approach to the co-ordination of behaviour in multi-agent systems. In this paper we examine the relationship between social laws and rational behaviour, by which we mean behaviour that can be justified by a defensible argument. We describe how social laws have previously been defined and used within the context of Action-Based Alternating Transition Systems (AATS). We then show how an account of argumentation for practical reasoning in agent systems, also based on AATS, can be used to determine what is rational for the agents to do in the absence and presence of such laws. The reasoning involved is both of a practical and epistemic nature: agents need to make decisions about what to do based upon the assumptions that they make about the states they find themselves in, and crucially, they also need to reason about what the other agents in the scenario will do. What is rational for the agents to do has implications for the need for social laws, the ways in which social laws can help the situation, the form the social laws should take, and the likelihood of compliance with the social laws. This paper demonstrates how we can think about social laws and rational behaviour in a single framework, so as to identify these implications in particular scenarios, and so frame social laws accordingly.

1 Introduction

Co-ordination within multi agent systems can be addressed through numerous different approaches. One important approach is through the use of social laws (e.g. [10][9]) that constrain the behaviour of agents within a scenario so that compliance with the law ensures that either some particular undesirable state is avoided or that some desirable state is eventually reached. In practice the realisation of social laws takes a variety of forms, ranging from mere conventions of etiquette, through moral conventions, to laws which have legislative force. In [12] it has been shown that such laws can be effectively expressed and understood using Action-based Alternating Transition Systems (AATSs) and Alternating-time Temporal Logic (ATL). In the absence of such laws, however, agents will not behave arbitrarily. In some cases, it may be enough for agents to behave rationally to guarantee the desired outcomes. In others the laws may be essential to guide

the behaviour of the agents. In still others it may be rational for one or more agents to violate the laws, potentially rendering them ineffective. Thus the need for, the benefits of, the form of, and the effectiveness of, social laws all require some consideration of what is rational for agents to do in the various situations, both where there are social laws and in the "state of nature" without them. In this paper we demonstrate how an argumentation based approach to practical reasoning, also based on AATS, can be used to determine what is rational for the agents to do in the absence of hard constraints enforcing such social laws. In doing so we consider a particular example, taken from [12], concerning the co-ordination of the movement of two trains. Using this example we consider how the reasoning differs depending upon the view that other agents take of their counterparts within the scenario, both what these other agents are likely to do, and the degree to which the interests of the other agents are respected. Our approach differs from that of Castelfranchi [7], in which social action is considered in terms of agents adopting the goals of others. On our view agents do not adopt goals of other agents, although their actions may further these goals, but rather choose to constrain their actions so as to enhance, or at least not threaten, the interests of other agents. Another characterisation of selfish and social agents is given in the context of the BOID architecture [6]. There agents are considered to have obligations as well as the standard beliefs, desires and intentions, and a selfish agent is one which prefers its desires to its obligations, and a social agent one which prefers its obligations to its desires. In our approach these conflicts are not resolved by a policy of this sort, but by considering the effect of the interests of the agents concerned: selfish agents are distinguished by preferring their own interests to even important interests of others.

The rest of the paper is structured as follows. In Section 2 we provide the details of the example scenario that we will use, which is taken from [12]. In Section 3 we briefly describe the background theory of practical reasoning that we use to enable agents to formulate and critique arguments about what to do, and so provide justifications of their actions. In Section 4 we show how this theory can be used to drive the reasoning in the scenario. In Section 5 we provide a general discussion of social laws that draws on our examples and covers consideration of when they are required, how they operate and what form they should take. Section 6 finishes the paper with some concluding remarks.

2 Social Laws

In [12] van der Hoek et al. make use of Action-Based Alternating Transition Systems (AATS) to explore social laws as a means of co-ordinating multi-agent systems, and we will use their notation. In an AATS transitions between states are governed by joint actions which are composed from the individual actions of the agents involved. A formal definition of AATS is given by van der Hoek et al. in [12]. It is their example that we will make use of here to extend the exploration to encompass consideration of what is rational for the agents to do in various scenarios.

The example has two trains, one running eastwards and one running westwards. For most of the circuit each train has its own track, but this narrows to a single track shared by the trains where the track enters a narrow tunnel. If both trains enter the tunnel together, therefore, they will crash. The trains may be in one of three states, *away* from the tunnel, *waiting* to enter the tunnel, or *in* the tunnel. At each point they may move (away to waiting to in to away) or stay still. Two particular aspects of the scenario are relevant: *safety* i.e. ensuring that there is no crash, and *progress* i.e. ensuring that the trains keep moving as much as possible whilst avoiding a crash. Initially they are both away from the tunnel. The transitions of the AATS for the scenario are shown in Table 1. Here the nine states in the scenario are labelled q0–q8. In each state each agent may choose one of two actions, move or do nothing. When the actions of the two agents are combined, this results in four joint actions: j0, where both trains do nothing; j1, where the eastbound train does nothing but the westbound train moves; j2, where the eastbound train moves but the westbound train does nothing; and j3, where both trains move. The final column of the table gives the interpretation function that shows which propositions are true in each state (away, waiting or in) subscripted for each train.

Table 1. Transitions/Pre-conditions/Interpretation

q/j	j_0	j_1	j_2	j_3	$\pi(q)$
q_0	q_0	q_1	q_3	q_5	$\{away_E, away_W\}$
q_1	q_1	q_2	q_5	q_6	$\{away_E, waiting_W\}$
q_2	q_2	q_0	q_6	q_3	$\{away_E, in_W\}$
q_3	q_3	q_5	q_4	q_7	$\{waiting_E, away_W\}$
q_4	q_4	q_7	q_0	q_1	$\{in_E, away_W\}$
q_5	q_5	q_6	q_7	q_8	$\{waiting_E, waiting_W\}$
q_6	q_6	q_3	q_8	q_4	$\{waiting_E, in_W\}$
q_7	q_7	q_8	q_1	q_2	$\{in_E, waiting_W\}$
q_8	q_8	–	–	--	$\{in_E, in_W\}$

The transitions can be shown diagramatically, as in Figure 1 below. In the diagram each of the states is labelled with its number (one of q0–q8) in the bottom right hand corner. Each state also contains two propositions to determine the status of each train: the top proposition represents the eastbound train's status, the bottom proposition the westbound's. Each proposition can be set to either 0, when the train is away, 1, when the train is waiting to enter the tunnel, or 2, when the train is in the tunnel. The arcs are labelled with the joint actions, as described in Table 1.

In general, the social laws approach is intended to constrain the behaviour of agents in particular states, so as to achieve certain objectives, typically that some state is avoided, or that some state is eventually reached. A social law is

said to be *effective* if compliance with the law ensures that the objectives are achieved. The main objective of the above example is to ensure that there is no collision, that is, that state q8 is never reached, with a secondary objective that the trains are able to make progress and so able to reach any of the states away, in and waiting. The undesirable state q8 can only be reached from one of the three states q5, q6 and q7. So, to ensure that q8 is not reached, a social law is needed to restrict the behaviour of the agents in these three cases. One such law proposed and formalised in [12], which we will call SL1, is as follows:

1. when both trains are waiting (q5) the eastbound train should not move;
2. when the westbound train is in the tunnel and the eastbound is waiting (q6) the eastbound train should not move;
3. when the eastbound train is in the tunnel and the westbound is waiting (q7), then the westbound train should not move.

As is shown in [12] this social law is effective, in the sense that if it is obeyed, it will ensure that the trains do not collide. As noted there, however, this law is asymmetric, in that it favours the westbound train over the eastbound train (although a very similar social law could be made which favoured the eastbound train, by modifying the first condition). Note that SL1 does not guarantee that any progress will be made since a train could remain in the tunnel indefinitely without violating SL1. SL1 thus seems to assume that agents will choose to move from the tunnel at the first opportunity, and so no social law is required to ensure that they do so. We will also make this assumption and focus our subsequent discussion on state q5. Of course, we could make a social law to ensure that this assumption is satisfied, consistent with the conditions of SL1.

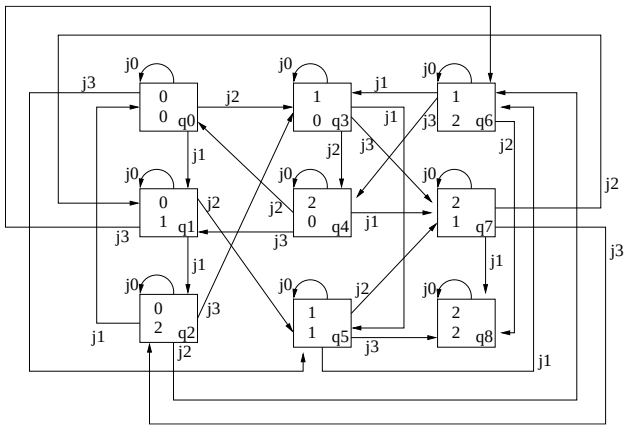

Figure 1: State transition diagram for scenario

The argumentation based model of practical reasoning that we make use of in this paper enables us to evaluate the law in terms of what the agents would choose to do in the absence of any social law. So, we view the situation as a practical reasoning problem. In our account there is no need for the agents to

reason about each others' beliefs since the relevant situation is fully described in the publicly available structure of the AATS. Moreover, in our example, agents have perfect information as to the situation and so disagreements as to how it is represented in the AATS do not arise. For this reason we need not make use of epistemic logic.

In the next section we provide an overview of the argumentation-based approach that we make use of to model the problem.

3 Background Theory of Practical Reasoning

In [2] an argument scheme and associated critical questions are presented to enable agents to propose, attack and defend justifications for action. Such an argument scheme follows Walton [13] 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, which can be used to overturn the presumption. The argument scheme AS1 developed by Atkinson [2] is an extension of Walton's *sufficient condition scheme for practical reasoning* [13]. AS1 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 Walton's notion of a goal has been made more precise by distinguishing three elements it encompasses: the state of affairs brought about by the action; the goal proper (the desired features in that state of affairs); and the value (the reason why those features are desirable)[1].

Agents act so as to bring about states of affairs that promote the particular values that are of concern to the individual agents. Thus, each agent has a preference ordering on the values it considers relevant in the particular scenario. We can therefore characterise agents' behaviour with respect to the ordering that they place on values. As mentioned previously, the two values of concern in the train scenario are 'safety' and 'progress'. *Prudent* agents will place a higher value on safety, but *reckless* agents will value progress more highly. In the examples that we discuss in Section 4 we use the terms *selfish* and *moral* to describe the behaviour of agents, following [3]. Note that when a value is promoted it is done so in virtue of one agent or both agents. Thus progress is promoted in virtue of the eastbound train moving, the westbound train moving, or both moving. Similarly safety is demoted when the eastbound train crashes, the westbound train crashes, or both crash. We therefore subscript values to show which agent

[1] In this sense values represent the social interests promoted through achieving the goal. Thus they are qualitative, as opposed to quantitative, measures of the desirability of a goal.

is progressing or safe. Selfish agents prefer their own interests to those of others and thus they will rank promotion of values in respect of themselves more highly than any values promoted in respect of others. For example, considering the value ordering from the perspective of a selfish prudent eastbound agent will give us the following: $\text{safety}_E > \text{progress}_E > \text{safety}_W > \text{progress}_W$. Moral agents, on the other hand, will take the other agents' interests into account, and so a prudent moral agent will believe that safety in respect of others is more important than its own progress. A truly moral agent will give the value equal rank, whichever agent is affected, but, as discussed in [3], it remains morally acceptable to prefer promotion of a value in respect of oneself to promotion of that value in respect of others, provided the ordering of values is consistent. Note that moral agents are not *sacrificial* (i.e. they do not favour promoting a value in respect of the other agent over promoting it in respect of themselves). Note also that the values are ordered according to the preference of the agent itself: a prudent agent is not required to rank the progress of a reckless agent more highly than that agent's safety, even though that is the preference of the other agent. Nor is a moral agent required to adopt values of the other agent if it does not recognise their worth: if the westbound train had a value "excitement" promoted by near collisions, the eastbound train would be under no moral compulsion to consider that value. Thus, considering the value ordering from the perspective of a moral prudent eastbound agent will give us the following: $(\text{safety}_E = \text{safety}_W) > (\text{progress}_E = \text{progress}_W)$.

Associated with AS1 are seventeen different critical questions [4] that challenge the presumptions in instantiations of AS1. 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, and their interpretation in terms of AATS, can be found in [4]. In the next section we make use of a selection of these critical questions and upon doing so we will make clear the attack that the critical question is asserting.

Using argument scheme AS1 and its associated critical questions to produce arguments for reasoning about matters 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. In the example that we provide in the next section the argumentation is expressed in natural language terms for ease of understanding, though we note that the machinery to express these arguments formally for use with an AATS is given in [4] and it would be a simple task to express them in this notation.

Finally, we note that the argument scheme AS1 can also be used in a negative form (AS2): given a particular set of circumstances, an action should not be performed, as it would lead to a particular state of affairs that entails some

'goal' which demotes a value. This negative version of AS1 can thus be used in scenarios where the onus is on avoiding some undesirable outcome rather than achieving some positive outcome, as is the case in the examples we present in the next section.

4 Example

Using the example scenario described in Section 2, we now show how the two agents in this scenario, the eastbound train and the westbound train, will each reason about what to do, in the absence of constraints, by instantiating the argument schemes AS1 and AS2, and posing the appropriate critical questions. We do so in respect of a number of different scenarios, based on the different possible value orderings that determine the agents' behaviour. We make the assumption in each case in this scenario that perfect information is available to both trains so that there is no epistemic uncertainty as to the current state. Now, in order to decide what to do, each agent will need to consider how it can promote its values, taking into consideration what the other agent will do. Of course, not all agents will act in the same way in a given scenario, and this will be reflected in the weight given to the arguments. We consider a number of different scenarios in turn, though the objective in every case is to avoid collision, i.e. avoid state q8.

4.1 Scenario 1: Reasoning in the Absence of Social Laws

We begin by considering how agents in the scenario will act in the absence of any social law. The reasoning starts in the initial state q0 where both trains are away. In q0 there are no controversial decisions to be made, each train may move knowing that the other is away. Each agent can thus instantiate an argument for moving, and can assume that the other agent will reason in the same way. However, since it is not possible to reach the state in which collision occurs from q0, the action of the other agent will not make an important difference and so need not be of real concern. Thus, the eastbound train will instantiate an argument as follows:

Arg1: In state q0, I should perform j2, to achieve waiting status in q3, promoting the value progress.

No attack can be successfully used against this argument since even if the assumption that the other agent will also choose to move is proved incorrect, the resulting state is equally good as far the eastbound train is concerned. We can see that Arg1 will similarly hold for states q1–q4 since none of the actions that can be performed in these states lead to the undesirable q8, regardless of how the other agent acts. Thus in each of these cases progress can be pursued without risk. The remaining states, q5–q7, can however, lead to q8. This is shown in Figure 2 below, which gives the subset of the scenario containing the states from which a collision can occur.

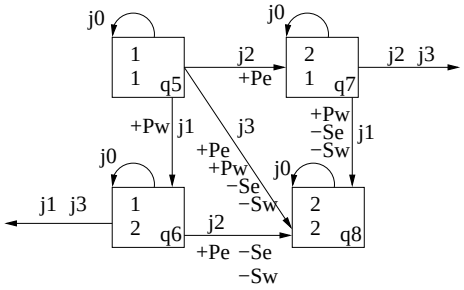

Figure 2: State transition diagram for states q5–q8

The arrows from states q6 and q7 that do not lead to any states signify that if these actions are performed, they will lead back to safe states (which are not of concern for this part of the scenario). Additionally, as in [3], in this diagram the transitions are labelled with the relevant values that are promoted or demoted by the transitions, with respect to each agent. Recall, the two values of concern are progress (P) and safety (S). Thus, for example, where a transition is labelled by +Pe, this indicates that progress is promoted for the eastbound train. For reasons of space we omit value labels from transitions that are neutral with respect to value promotion (i.e. where the transition neither promotes nor demotes safety or progress for either agent).

Selfish Agents Let us consider first the state q5 from the perspective of the eastbound train using the assumption that both agents are selfish. Since selfishness means that the agent will want to better its own interests, it will instantiate an argument to move in order to promote progress. When there is more than one action that will promote progress, the agent will choose the one (if any) that does not demote some other value. So, considering the eastbound train in q5, the following argument will be put forward:

Arg2: In state q5, I should perform j2, to reach q7 and enter tunnel, promoting the value progress.

We can immediately see that this argument poses a problem if the other agent, reasoning in the same way, also decides to move, since this will result in j3, rather than j2 being performed, which leads to the collision in q8. We can thus pose a critical question against this instantiation. The particular critical questions that is applicable here is CQ17: *can the other agent be guaranteed to perform its part of the joint action?*[2]

Obj1: Agent W cannot be guaranteed to act so as to execute j2.

Remembering that each agent in this scenario expects the other to act in a selfish way and thus move, we can see that the objection raised in this critical question is upheld; the agent can be expected to act in this way since, being selfish, it will want to promote its progress by moving, so j3 will be executed.

[2] Arguments that instantiate a critical question are labelled with 'Obj', to distinguish them from those arguments that instantiate an argument scheme ('Args').

Thus, Arg2 is defeated by the attack of Obj1 and will be abandoned. Now the eastbound agent must consider whether there is any argument for executing j3:

Arg3: In state q5, I should perform j3, to reach q8 and enter tunnel, promoting the value progress.

Whilst we can see that executing j3 will indeed promote progress, it will, however, lead to the state in which collision occurs. Thus we can critically question Arg3, since the action has a side effect that demotes another value. CQ9 raises such an objection:

Obj2: Action j3 has a side effect that demotes the value safety.

Again, we can see that this objection is upheld and since normally safety is more important than progress, Arg3 will be abandoned. This leaves only one choice for the eastbound train: to stay still. The agent will be indifferent as to whether j0 or j1 is executed since both have the same effect with respect to value promotion, each being neutral with respect to both its values. Whilst j1 does in fact promote progress for the westbound train, this has no influence on the eastbound train since it reasons in a selfish manner. If we consider the possibility that the eastbound train would choose to execute j0, this argument would again be subject to questioning through CQ17. In response to this we can say that the westbound train, being selfish, will actually prefer to move and so state q6 will be reached with no detriment. This would bring the reasoning round full circle since the westbound train would then need to consider whether the other agent would act so as to guarantee that j1 would be executed. We can thus see that that a symmetric set of arguments to those given above would be generated by the westbound train. The overall result of the reasoning would mean that each train would remain still, as shown through the following argument:

Arg4: In state q5, I should perform j0, which would avoid collision, so as not to demote the value safety.

No critical question can be successfully posed against this argument. However, again, although the collision is avoided for prudent agents, the reasoning results in an undesirable situation since deadlock is created as neither train has an argument to move. It is from the need to avoid this effect that the requirement for a social law arises. In the absence of such a law the deadlock is broken only when one of the agents becomes sufficiently reckless to prefer progress to safety. Should both agents do so at the same time, a collision will occur.

So far we have assumed that the agents in this scenario are all acting selfishly. We should therefore consider if the outcome of the reasoning would be any different if the agents are not in fact selfish, but instead 'moral' agents. That is, they are not selfish, but neither are they sacrificial in that they do not favour the other agent's interests over their own.

Moral Agents Starting over where the initial state is q0, there are again no potentially dangerous actions so each agent will choose to move. The problem states remain q5–q7. Again in q5, each agent will have an argument to move, as in the previous scenario, and Arg2 will be put forward. Likewise, CQ17 can

again be posed to state that the other agent cannot be guaranteed to act so as to execute j2. However, this time when the eastbound train considers that this non-compliance with j2 will lead to j3 being executed and subsequently q8 will be reached, it will, unlike in the previous scenario, consider the values of the other agent. So, whilst the same line of reasoning will still apply, the eastbound train will now have an *additional* attack, through the use of CQ9, that can be posed against Arg3:

Obj3: Action j3 has a side effect that demotes the value safety of the westbound train.

In order to see how this extra argument affects the evaluation of the set of arguments, we can organise them into a value-based argumentation framework (VAF) [5], which is an extension to Dung's abstract Argumentation Frameworks (AFs) [8]. AFs provide a means of evaluating the acceptability of a set of arguments in terms of the attack relations between them. VAFs extend Dung's AFs to accommodate different *audiences* with different values and interests. Within a VAF, which arguments are accepted depends on the ranking that the audience (characterised by a particular preference ordering on the values) to which they are addressed gives to the values motivating the argument. In essence, attacks are removed if the value of the attacking argument is ranked below the value of the attacked argument. The VAF for the arguments relevant to this scenario, from the viewpoint of the eastbound train, is given in Figure 3.

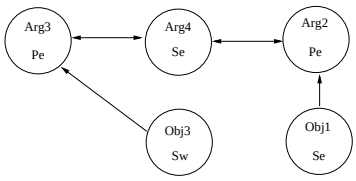

Figure 3: VAF for scenario 2.

To evaluate the status of the arguments we need to consider the preference ordering on values to resolve the conflicts between the arguments. As stated previously, safety is preferred to progress in all cases, thus Arg3 will always be defeated. The point to note however, is that the attack of Obj3 on Arg3 succeeds in this scenario since agents consider each other's values. But, this particular attack would not succeed for the previous case where the agents are all selfish, even though Arg3 would still be defeated by Arg4, assuming the agents are prudent. The result of the reasoning here is that both trains will again remain still, but this time there is an additional reason, which will prevent even a reckless agent from moving, provided it is moral enough to consider the other agent's safety, even if it does not care about its own. Thus, although the reasoning on the parts of both agents will avoid the undesirable q8, here a social law is needed, not to avoid collision, but in order to break the deadlock. In this case, where the agents are considerate of each other in this way, the need for a social law is greater, since even recklessness will not help.

The scenarios discussed so far consider a situation which was symmetrical with respect to the interests of the agents. Now suppose we alter the scenario

so that one of the agents, say the westbound, is instead a pedestrian and both agents are again assumed to act selfishly. In this case, the state to avoid remains q8, but the reasons for avoiding it have now altered since a collision would only impact negatively upon the pedestrian (where 'safety' and 'progress' remain the only two values of concern): the train would be unharmed by the collision.

Considering again the problematic q5, the reasoning of the eastbound agent will again begin with the proposal of Arg2, against which CQ17 can be posed. In this case the train's safety is not now compromised in q8: the transition between q5 and q8 will be labelled only with the value 'safety' demoted in respect of the pedestrian. Thus, the eastbound train is aware that the pedestrian, not willing to compromise his safety, will not move, leaving the train free to enter the tunnel. But, now the critical question CQ9 does not apply, since even if the pedestrian ignores the risk, the resulting situation will not demote the safety of the train. In this case the rationality of the agents should be enough to ensure that q8 is avoided and the need for a social law to enforce such behaviour is dispelled, since the agents will behave in compliance with SL1 anyway. However, we may wish to actually enforce such behaviour through the issue of a social law since consideration must be given to the situation where an agent may be reckless and a collision ensues. The law now, however, is for the pedestrian's own good, rather than to provide the co-ordination required in the case of two trains. In such a case we may wish to implement punishments or sanctions against such behaviour, as we discuss later in Scenario 3.

The previous example considers the reasoning of the agents where one is a pedestrian, the other a train and both are selfish. Does the outcome change if the agents are moral? Again, considering the problematic state q5, we can see that the reasoning of the train will begin by it proposing Arg2, against which CQ17 can again be posed. Once more, the train would expect the pedestrian to comply with not entering the tunnel, yet in this case the train will act in consideration of the pedestrian's values in addition to its own. Thus, the VAF for the situation will be updated so that the argument based on the eastbound train's safety no longer appears in it. Here the train will not move since the argument for moving is defeated by the argument demoting the pedestrian's safety. However, the pedestrian, also reasoning that the danger only concerns himself, will not move either, leading to a deadlock situation. In this case a social law is clearly needed to avoid deadlock, even if only the train is moral.

4.2 Scenario 2: Reasoning in the Presence of Social Laws

We now consider how the reasoning will change when a social law is present. As demonstrated above, reasoning in the scenario in the absence of a social law will indeed avoid the undesirable state, but a deadlock situation arises. This is true for both the case in which the agents are selfish, and that in which they act in accordance with 'morality'. Now let us consider the effect of introducing the social law SL1, as stated in Section 2. When such a social law is in place, the agents in the scenario have a change in information about the actions of each other. Thus all agents may still act selfishly, i.e. in accordance with their

own interests, but there is now an assumption that the other agents will all obey the law. This effectively excludes certain joint actions (those containing the prohibited action) from the AATS. So, for our problematic situation, state q5, the social law ensures that the agents will not act so as to end up in q8, and the deadlock is broken through the law specifying which agent should move into the tunnel first. This means that the westbound train will generate an argument for moving:

Arg5: In state q5, I should perform j1, to reach q7 and enter tunnel, promoting the value progress.

CQ17 can still be posed against this argument to test the presumption that the eastbound train can be guaranteed to act so as to execute j2. However, the response to this argument is now that the eastbound train will act so as to execute j2, because it will obey the social law. But, as noted previously, SL1 is asymmetric in that it favours the westbound train over the eastbound train. Nonetheless, even though the agents are not treated equally by the law, it does in fact provide more benefit to them both than the case where there is no law. Since the choices that are forced through the law are rational in any case, it follows that each train will move sooner, and without the need to degenerate into recklessness, than they would if the social law was not in place. In this way adherence to the law is reinforced through rationality. The same outcome is also true for the situation in which the agents are 'moral' as opposed to selfish.

If we now return to the example where one agent is a pedestrian and the other a train, we noted previously that a social law is required where the agents are acting morally. Again, we consider the application of SL1 in this situation. As before, the law works so as to remove the deadlock, but since a choice must be made as to which party will get to move first, the law will again favour one of the parties. However, in order to reinforce the behaviour that rationality suggests, the choice of who goes first in this case should not be an arbitrary one. Here, the social law should be defined so as to allow the train to move first, since it is the party against which no danger is posed. Now the moral train can enter the tunnel assured that the pedestrian will wait, and any threat to the pedestrian's safety comes from his own disobedience. If, on the other hand, the law tried to make the train wait, the train would have no reason other than conforming to the law to wait, and so there would be temptation to violate the law. Moreover, the pedestrian might be reluctant to jeopardise his safety by trusting that the train would comply. Such a law might therefore lead to collisions when the train disregarded the law, and to deadlocks when the pedestrian did not trust the train to comply. By reinforcing rather than conflicting with the rational choice, the law is more likely to be followed since it does not penalise the party who has a rational justification for non-compliance. This suggests that in general when framing the law we should consider which parties benefit the most when compared with the situations without the law.

4.3 Scenario 3: Social Laws with Sanctions

As noted above, the presence of social laws should prescribe[3] the behaviour of the agents. However, since the agents are autonomous, they cannot always be guaranteed to adhere to the social laws in place. Thus, we consider how obedience to the law can be achieved through the use of sanctions. Sanctions can take two forms; they may operate in relation to a value representing the stigma associated with violating the law, or they may operate through undesirable consequences relating to the state reached when the law is violated i.e. the agent is in some way punished for violating the law. In the presence of these new elements we now consider how the reasoning in the scenario will differ in state q5.

We begin with the case of the selfish agents. In q5 SL1 states that the eastbound train should remain stationary. However, this agent, as in the previous scenarios, will have an argument for moving into the tunnel based on pursuit of progress. There remains an argument against moving, but a reckless agent may be tempted to ignore this and move in an attempt to get into the tunnel first, thus violating the law. However, we can now introduce the third value of 'honour' into the scenario, whereby any transition that ignores the law and subsequently takes an agent from a 'safe' state into the undesirable q8, demotes this value. All actions that are executed in adherence with the law will promote the value. So, where there may be temptation to break the law, e.g. the eastbound train does not want to wait in q5, there is now another critical question, based upon the demotion of another value, that can be posed against the argument to move:

Obj4: Action j2 has a side effect that demotes the value honour.

In the simple scenario considered here it may be that honour does not feature highly in a value ordering and thus Arg3 will resist the attack of Obj4. However, in a more complex scenario where honour plays a role in future interactions, this may be enough of a sanction to deter violation. As an additional, or alternative, to this form of sanction we may also take a quantitative measure into account. We can thus add a proposition to each of the states to represent some kind of monetary possession. Here, where an agent violates the social law the state reached will actually be different to that intended through the application of the sanction, as expressed through critical question CQ2:

Obj5: Action j2, does not lead to q7.

Of course, with the addition of this proposition the state transition diagram will now need to be altered to show in which states money is decreased for each agent. So, the losses made in such situations should be enough to deter the agent from violating the law here, if money is ranked higher than progress. We can see that this is the case by considering the VAF for the arguments shown in Figure 4, where the values 'honour' (H) and 'money' (M) are introduced to ground the appropriate arguments.

[3] Although we make use of deontic notions such as obligations and their violations, we do not give any precise characterisation of them here. The relationship between deontic logic and ATL is the topic of [14] which introduces Normative ATL and provides definitions of obligation and permission in terms of an AATS.

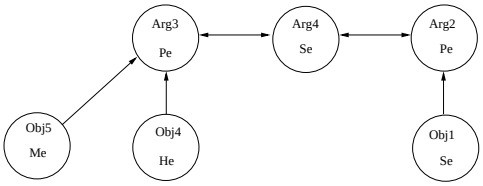

Figure 4: VAF for scenario with sanctions.

The above view is for that of selfish agents, so we again consider how the reasoning changes in the case of moral agents, for the scenario where there is a social law with sanctions. Here, unlike the case of selfish agents, the temptation to violate the law will be removed since the agents take each others' interests into account; an objection based on CQ9 can again be posed against Arg3 to state demotion of the other agent's safety, which will be enough to stop the violation.

However, we note that the above holds for this particular scenario because there are no conflicts *within* a value, i.e. in our scenario safety always trumps progress and there is no situation here where an agent is forced to choose between its own safety and the other agent's, nor its own progress and the other agent's. There are, however, other example scenarios where such a choice is required, and these in turn could lead to the temptation to violate a sanction. In such cases we need need to make a distinction between different levels of morality, in order to resolve the conflict. One such account of these different levels has been given in [3]. There, a distinction is made between 'moral' agents, as we have used in our example where values are ordered but within each value agents are treated equally, and 'noble' agents, where values are ordered in a moral sense, but within a value an agent prefers the other's interests. For example, a noble eastbound agent would order values as follows: $\text{safety}_W > \text{safety}_E > \text{progress}_W > \text{progress}_E$. There are numerous everyday examples that can be alluded to in which such a distinction is required. Consider the scenario of being sat on a train in which all the seats are occupied and a pregnant woman boards the train. In such a case, the norms of society are such that it is expected that a person with no impediment would give up their seat for the woman. Whilst a moral agent would value his own comfort equally to other peoples', he would be required to be noble, i.e. value the comfort of a pregnant woman over his own, in order to be forced to act and give up his seat. Whilst temptation to violate such a norm would not exist in the case of a noble agent, for the moral agent, whose value ordering conforms to a lesser standard of morality, the temptation would arise. Here, a sanction based upon demotion of honour could again be employed in order to force compliance.

Finally, concerning sanctions, there is one further problem to be considered before too much reliance is placed on them: sanctions require that the transgression be detected and the punishment enforced. But this will not always be the case. Sanctions therefore require an additional agent, the agent responsible for enforcing the social laws, to be modelled in the system, and additional joint actions since the sanction may or may not be enforced. In many cases, this reintroduces uncertainty into the situation, since the agent cannot know that

the transgression will be followed by the sanction. So an agent reasoning in the presence of sanctions will still have an argument for the transgressive act, albeit one subject to CQ17, since the sanction may be enforced and the expected state not reached. The agent's decision will then need to balance the risk of detection against the gains resulting from transgression. In such cases other agents similarly will not be as sure of their assumption that the law will be complied with, since conformity now requires a degree of judgment on the part of the other agent. In many cases therefore, where detection is not assured, the social laws fail to provide the essential increase in certainty as to how others will behave.

5 Discussion

From the above considerations, we can attempt to draw some generalisations, about when social laws are required, how they operate and the form they should take. We will draw upon the above discussions, and additionally add some illustrations from road traffic practice.

In some cases no social law is really needed, as illustrated above where the social law requires the pedestrian to give way to the train, since the agents will avoid the collision, provided they act in their own interest and expect the other agent to do so. This is the situation with regard to pedestrians crossing the road: since prudence should lead them to avoid crossing in front of cars, no law is necessary, although parents do try to teach their children to value safety over progress.

In other cases, represented in the example by the scenario with two trains, a social law is needed, since agents will be prevented from acting freely in accordance with their preferences because of uncertainty about what the other agent will do. What the social law does is to remove this uncertainty: since the social law requires the eastbound train to give way, the westbound train can confidently enter the tunnel. In road traffic scenarios, this gives rise to conventions as to which side of the road should be driven on: it is crucial that the agents know what the others will do if collisions are to be avoided. Of course, such a convention does constrain behaviour but agents will willingly accept the constraint since everyone gains: the preference is to drive on the same side as others, not on a particular side. While in this case the effect of the law is the same for both agents, in the train case the westbound train gains more, since it is effectively enabled to act in accordance with its best interests without fear that the eastbound train will act so as to endanger it. Nonetheless, the eastbound train also gains, since the social law assures it of eventual progress without danger, whereas, without the social law progress cannot be made without becoming reckless. The role of the social law here is purely one of *co-ordination*.

In other situations agents are constrained with no advantage to themselves. The example scenario is where the train is required to give way to the pedestrian. Since the train is in no danger from the collision, the arguments to act in conformity with the law relate to benefits to the pedestrian, not the agent which is constrained. On roads this is the situation with pedestrian crossings: it

is concern for the safety of the pedestrian that should induce the car to comply with the convention that the pedestrian has right of way in such cases. The social law is socially justified in that it gives a substantial benefit to one agent at a small cost to the other, but compliance does rely on a degree of moral sense on the part of the car drivers.

Social laws essentially work by reducing uncertainty as to what others will do in a given situation, thus allowing the consequences of one's own actions to be more predictable. Sometimes there will be gains for everyone, so that selfish agents will, given the assurance about how others will act, rationally conform. In other cases it may be necessary for the agents to consider the values of others to motivate conformity. This is illustrated by advertisements encouraging conformity to speed limits which emphasise the danger to oneself in the case of motorways, and the danger to others in the case of low speed limits in residential areas. The former can appeal to self interest, but the latter requires a sensitivity to the interests of others. In some extreme cases – military draft in wartime may be an example -- conformity requires the agent to put the interests of others before of its own most valued interests, and here sanctions will be essential to produce conformity.

In her work on emergence of norms, Ullmann-Margalit [11] distinguishes norms of co-ordination, where both parties benefit from compliance, from norms of co-operation where an increase in the common good comes at the price of a decrease of individual goods. This is the situation represented by the classic game theory scenario known as the Prisoner's Dilemma, and she refers to such norms as PD-norms.

Co-ordination norms are unproblematic, since given the resolution of uncertainties offered by the social law, rational agents will freely choose to comply with them. For PD-norms, however, the rational situation is defection, not compliance, as is well established in game theory. Ullmann-Margalit suggests three ways of inducing compliance: making defection impossible, making defection unattractive through sanctions, and what she calls "honour", which involves a sufficiently strong sense of identity with the other agent to mean that the interests of the other are given sufficient weight to induce co-operation.

The first method, making violation impossible, is the approach typically taken in Electronic Institutions e.g. [1]. There, for example, if a participant in an auction is not permitted to bid according to the norms of the institution, this action is simply unavailable. While this is possible in a structured situation such as is provided by an electronic institution, there are several problems with this as a general solution. First it violates the autonomy of the agents: they are forced to obey the norm, and so their freedom is constrained. Secondly agent interaction in open systems is desirable in less structured contexts, when it is impossible to impose these constraints. But most importantly, it is part of the nature of social laws that there are occasions when it is desirable that they are violated. Occasionally it is necessary to drive on the wrong side of the road to avoid an accident: we would not want this to be impossible. In a medical emergency we may not only allow, but desire, speed limits to be exceeded. In complex environ-

ments norms can conflict, and we would wish our agents to solve this conflict rationally with regard to the particular situation, rather than blindly following the norms.

Sanctions, as mentioned above, can be effective, given a regime in which detection is sufficiently certain and the punishments sufficiently great. This requirement, however, may be very difficult to achieve in a loosely structured environment. In an agent society, however, there are further problems: what sanctions are appropriate to agents, and how can they be applied? Possibly the best that can be done is through honour, but the opportunities to cloak identities in cyberspace make this at best a flimsy defence.

For the third strategy to be possible we need to have agents that have the ability to reason about what to do in a particular way, so that they consider the general interest as well as their own. In so far as PD-norms are desirable in situations where neither making violation impossible nor enforcing sanctions is practicable, this seems to be the only solution. Respect for social laws and consideration for others are necessary parts of the functioning of human society: without a certain degree of compliance with social laws through simple consideration of others, life would be intolerable. It might be thought that consideration for others should be a desirable feature of agent societies also: we would not employ a person we believed to be amoral or dishonest, so why should we be prepared to unleash agents with no sense of moral duty on an unsuspecting world? We bring up our children to respect the interests of others, so should we not implement our agents in the same way?

6 Concluding Remarks

In this paper we have considered social laws in the context of rational decision making. We have seen that at one extreme some social laws simply provide the necessary degree of certainty about how others will behave to enable good results to come from rational, self interested action. At the other extreme, other social laws will require the backing of certain and heavy sanctions to make compliance rational. In between there are situations where rationality leads to compliance if the welfare of the other agents involved in the situation is taken into account. When framing social laws we need to consider whether they will be adhered to: doubt as to the compliance of other agents will restore the uncertainty the social law was designed to resolve. When framing social laws, these factors need to be considered: sometimes that will lead us to prefer one formulation over another. There are also implications for designing reasoning agents: social laws will often depend on some sense of social obligation for their effectiveness, and so agents need to be designed to be capable of reasoning so as to consider the interests of others.

References

1. J. L. Arcos, M. Esteva, P. Noriega, J. A. Rodrguez, and C. Sierra. Engineering open environments with electronic institutions. *Journal on Engineering Applications of Artificial Intelligence*, 18(2):191–204, 2005.
2. K. Atkinson. *What Should We Do?: Computational Representation of Persuasive Argument in Practical Reasoning*. PhD thesis, Department of Computer Science, University of Liverpool, Liverpool, UK, 2005.
3. K. Atkinson and T. Bench-Capon. Addressing moral problems through practical reasoning. In L. Goble and J.-J. C. Meyer, editors, *Deontic Logic and Artificial Normative Systems*, LNAI 4048, pages 8–23. Springer, 2006.
4. K. Atkinson and T. Bench-Capon. Action-based alternating transitions systems for arguments about action. In *Proceedings of AAAI 2007*, pages 24–29, 2007.
5. T. Bench-Capon. Persuasion in practical argument using value based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–48, 2003.
6. J. Broersen, M. Dastani, J. Hulstijn, Z. Huang, and L. van der Torre. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *Proceedings of Autonomous Agents 2001*, pages 9–16. ACM Press, 2001.
7. C. Castelfranchi. Modelling social action for AI agents. *Artificial Intelligence*, 103(1-2):157–182, 1998.
8. 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.
9. Y. Moses and M. Tennenholtz. Artificial social systems. *Computers and Artificial Intelligence*, 14(6):533–562, 1995.
10. Y. Shoham and M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies. In *Proceedings of AAAI 1992*, pages 276–281, 1992.
11. E. Ullmann-Margalit. *The Emergence of Norms*. Clarendon Press, Oxford, 1977.
12. W. van der Hoek, M. Roberts, and M. Wooldridge. Social laws in alternating time: effectiveness, feasibility and synthesis. *Synthese*, 156(1):1–19, 2007.
13. D. N. Walton. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1996.
14. 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.