Rules are Made to be Broken

Trevor Bench-Capon[1] and **Sanjay Modgil**[2]

Abstract. There is an increasing need for norms to be embedded in technology as the widespread deployment of applications such as autonomous driving and warfare and big data analysis for crime fighting and counter-terrorism becomes ever closer. Current approaches to norms in multi-agent systems tend either to simply make prohibited actions unavailable, or to provide a set of rules (principles) which the agent is obliged to follow, either as part of its design or to avoid sanctions and punishments. We argue that both these approaches are inadequate: in order to meet unexpected situations agents must be capable of violating norms, when it is appropriate to do so, either accepting the sanction as a reasonable price to pay, or expecting the sanction to not be applied in the special circumstances. This in turn requires that agents be able to reason about what they should do from first principles, and one way to achieve this is to conduct value based reasoning using an argumentation scheme designed for practical reasoning. Such reasoning requires that agents have an acceptable set of values and an acceptable ordering on them. We discuss what might count as an acceptable ordering on values, and how such an ordering might be determined.

1 Introduction

As noted in the workshop call for papers, there is an increasing need for norms to be embedded in technology as the widespread deployment of applications such as autonomous driving and warfare and big data analysis for crime fighting and counter-terrorism becomes ever closer. Current approaches to norms in multi-agent systems tend either to simply make prohibited actions unavailable (e.g. [33]) or to provide a set of rules (principles) which the agent is obliged to follow, in the manner of Asimov's Three Laws of Robotics [4]. Neither of these methods can be seen as satisfactory ways of providing moral agents (i.e agents able to reason and act in accordance with norms) since not only is it in the nature of norms that they *can* be violated, but circumstances may arise where they *should* be violated. In fact norms are, in real life and also in MAS, typically backed by sanctions [10]. The idea behind sanctions is to change the consequences of actions so as to make compliance more pleasant and/or violation less pleasant[3]. As noted in [10], sanctions can be seen as *compensation* (like library fines) when they can be viewed as a charge for violation, which makes the situation acceptable to the norm issuer, or as *deterrents*, where the sanctions are meant to ensure compliance by relying on the self-interest of the norm subject. When the norm *should* be violated sanctions may be problematic as they disincentivise the agent. This problem can be lessened in cases where the violation can be condoned and the sanction not applied, but this requires an agreement between the agent and the agent imposing the sanction that the violation was justified (often not the case: consider dissidents such as Gandhi and Mandela). Moreover sanctions need to be enforced, otherwise agents may take the risk of escaping punishment, and violate the norm when there is no acceptable reason to do so.

Thus an important reason for thinking in terms of norms is the recognition that on occasion they need to be violated [24]. While the norm is intended to provide a useful heuristic to guide behaviour, allowing for a quick unthinking response, unreflecting adherence to such moral guidelines is not what we we expect from a genuinely moral reasoner. R.M. Hare, a leading moral philosopher of the last century, expressed it thus [22]:

> 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.42)

So while principles may serve well enough most of the time, there are situations where we need to think through the situation from scratch. In this paper we will consider how we can give software agents the capacity to perform quasi-moral reasoning[4].

2 Problems With Current Treatments

There are two main approaches to enforcing normative behaviour in MAS: either by removing prohibited actions (e.g. [33]), or by including explicit rules expressing the norms, often accompanied by

[1] Department of Computer Science, University of Liverpool, email: tbc@csc.liv.ac.uk
[2] Department of Informatics, King's College, London
[3] In decision theoretic terms, the ideal for deterrence being for violations to yield an overall negative utility.

[4] We say "quasi-moral" since software agents do not themselves have ethical status, and cannot be considered to share our values. In this paper we will see such agents as proxies for human beings in simulations or transactions, and so their values will be those of the human they are representing. Developing a set of values applicable to software agents would be the topic of another paper. To see that human values are not applicable to software agents consider the fact that their life is of little value, since they can be easily reproduced or replaced, they don't feel pleasure or pain, nor happiness nor sorrow, and have no experience of liberty or fraternity.

sanctions. Neither are entirely satisfactory. We will illustrate our discussion with a model of the fable of **the** *Ant and the Grasshopper* [1], previously used in [14]. The model takes the form of an Alternating Action-Based Transition (AATS) [33], augmented with value labels [6]. The transition system, in which the nodes represent the states the agent may reach and the actions it may use to move between them (in an AATS they are *joint* actions, one action for each relevant agent), is a typical ingredient of Multi Agent Systems (MAS): the value labelling provides the basis for moral reasoning.

In the fable the ant works throughout the summer, while the grasshopper sings and plays and generally indulges herself. When winter comes and the ant has a store of food and the grasshopper does not, the grasshopper asks the ant for help. The ant refuses and says the grasshopper should have foreseen this, and so the grasshopper starves. The same model also can be used to represent the parable of *the Prodigal Son*, except that in the parable the father welcomes the repentant prodigal back, and does give him food.

Using the first approach we would enforce the behaviour recommended by the fable by removing the transition from q_6 to q_5 or the behaviour of the parable by removing the transition from q_6 to q_7. A real life example in which actions are made unavailable is erecting bollards to prevent vehicles from entering a park (to use the famous example of Hart [23]). What can be wrong with this approach? After all, we can *prove* that the undesirable situation will not be reached, either using model checking [17] or analytic methods. Thus we can prove that universal compliance with the norm will achieve the desired results. This may be so, so long as the situation envisaged in the model is in operation. But suppose some state not modelled arises: perhaps someone has a heart attack in the middle of the park and so it is essential for an ambulance to enter the park in order to save that person's life. Now the bollards will prevent the person from being saved, and the object of the norm, i.e. the value that the norm is designed to serve, the safety of park users, will be demoted rather than promoted. While the norm is effective in an ideal world, we do not live in an ideal world, and in a sub-ideal world it is often the case that adhering to the norms applicable to an ideal world will not lead to the most desirable results[5].

Similarly, principles may cease to prescribe the best course of action in unforeseen situations. The whole point of Asimov's three laws as a fictional device is that following them may lead to outcomes that the principles were designed to avoid. While any set of principles may provide good guidance most of the time, it is not difficult to think of gaps, situations and conflicts where following the principles will lead to undesirable results, and so need to be disregarded. The problem is not improved by the existence of sanctions, and indeed may be made worse since the threat of possible punishment makes violation less attractive to the agent.

Thus while either of the approaches may be effective in closed systems (providing they are simple enough for a model covering every eventuality to be constructed), they cannot be sure to cope with the unexpected events and states that will arise in an open-system, where not every possibility can be envisaged or modelled[6]. In such cases we may find that the very reasons which led to the adoption of a norm will require the agent to violate that very same norm.

Irrespective of which option is chosen, the regulation of behaviours at the level of norms does not allow for agents to appropriately violate norms, in cases where compliance with the normatively prescribed behaviours results in demotion of the values that these norms are designed "to serve", or even of other, preferred, values. Hence, we argue that agents should be equipped with the capacity to reason about values, the extent to which normatively prescribed actions serve these values, which values are more important than other values (i.e. value orderings qua 'audiences'), and the ability to derive these orderings from a variety of sources, including experience, the law, and stories prevalent in the culture. These capacities constitute moral reasoning from first principles; the kind of reasoning required to deal with new and unexpected situations in which blind compliance with norms may lead to undesirable outcomes. This paper serves as a call to further develop reasoning of this kind, building on a number of existing developments that we survey.

[5] This is known in economics as the *Theory of the Second Best* [25].

[6] As Wilde put it in *An Ideal Husband*: "To expect the unexpected shows a thoroughly modern intellect".

3 Value-Based Reasoning

A method for value-based reasoning was proposed in [8], formalised using an AATS labelled with values in [6] and further articulated in [5], and which gave nine reasons for action in terms of the promotion and demotion of values. The basic idea is that the transitions which promote values form the basis of arguments for the action which will allow that transition to be followed, and that the transitions which demote values will supply arguments against actions which permit these transitions. Further arguments may come from assumptions about the current state and the state that will be reached by following a particular transition. These arguments and the attack relations between them (determined according to the so-called critical questions listed in [6]) define an argumentation framework [20]. Moreover since the arguments will be associated with values, the framework is a value-based argumentation framework (VAF) [9]. In a VAF the arguments are evaluated from the perspective of an *audience* (cf [31]) characterised as an ordering on values, and attacks which are unsuccessful for an audience are distinguished from those which succeed (*defeats*). The result is a set of arguments acceptable to a particular audience. If there are no cycles in a single value, this set will be non-empty and unique [9].

If we consider the ant's choice in q_6 of Figure 1, he may either refuse or give. Which is chosen will, using the labels of Figure 1, depend on whether the ant prefers his own pleasure to the life of the grasshopper. The application of value based reasoning to moral decisions was considered in [7], which suggested that moral acceptability required that one's own lesser values should not be more highly ranked than more important values relating to others. This would not (morally) allow the preference of the ant's pleasure over the grasshopper's life, and so require the ant to give food to the grasshopper. But the labelling in Figure 1 is not the only one possible. If we think more abstractly we may see the ant's refusal as promoting *Justice*, since the grasshopper knew full well that food would be required in the winter and not working in the summer would mean later exploitation of the good nature of the ant. Similarly we could label the giving of the food as *compassion* or *mercy*. Preferring justice to mercy becomes more legitimate if we consider the role of the moral code to be producing a sustainable society, which requires that working in the Summer be seen as the norm. As shown in [27] the sustainability of norms requires that transgressions be subject to punishment, and so punishing the grasshopper may be seen as the duty of the ant. Note too that in the parable the prodigal is repentant, and so the father will only be expected to show compassion once. Representing such things as repentance will require an extension to the state descriptions to record histories, but will allow a preference for justice over compassion to be dependent on the misbehavior being repeated. Benefits of tolerance of limited misbehaviour before en-

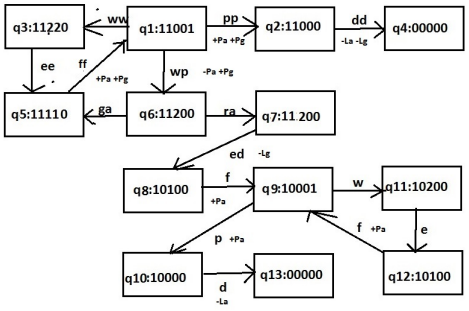

Figure 1. AATS+V: w = work, p = play, a = ask, g = give, r =refuse, e = eat, f = feast d =die. The same AATS+V is used for both the fable and the parable. Joint actions are ant/father, grasshopper/son. States are: ant/father alive, grasshopper/son alive, ant/father has food. grasshopper/son has food, summer/winter

forcing punishments is explored through simulation in [26].

Yet another way of describing the problem would be to recognise that the singing of the grasshopper may be a source of pleasure to the ant as well as to the grasshopper. Seen this way, the ant does not so much give food to the grasshopper as to pay for services rendered. This in turn requires requires recognition that it is the duty of the ant to pay for the services of the grasshopper, and so justice is now promoted by following the transition from q_6 to q_5, not q_7 . Moreover since a single grasshopper may entertain a whole colony of ants, the burden falling on a single ant may be relatively small.

If, however, there is only a single ant, suppose that the harvest fails, and there is no surplus to pay the grasshopper. Should the ant follow the norm, pay the grasshopper and starve or renege on the agreement and watch the grasshopper starve? Here we will have a genuine moral dilemma, in which the ant must choose between justice and its life. The ant may choose death before dishonour, but may also choose to renege with good authority. Thomas Aquinas writes:

> if the need be so manifest and urgent that it is evident that the present need must be remedied by whatever means be at hand (for instance when a person is in some imminent danger, and there is no other possible remedy), then it is lawful for a man to succor his own need by means of another's property, by taking it either openly or secretly: nor is this properly speaking theft or robbery.[7] [2], Question 66, Article 6.

Thus the ant has a choice, and either option can be justified. What the ant will do will depend on its value preferences. Arguably the original contract was foolhardy - on the part of both - since the failure of the harvest could have been foreseen by both parties, and whichever suffers has only themselves to blame.

4 What Makes a Moral Audience?

As the last example shows, there may be more than one morally acceptable ordering on values. Some other orderings, such as a refusal to pay the grasshopper even when there a surplus available to do so, are not acceptable. What we must do is to provide our agents with an acceptable ordering on which to base their reasoning. In order to do so, we need to look at the value order prevailing in society. As noted in work on AI and Law, the decisions made by courts often manifest an ordering on values. The case law decisions often turn on the value preferences the judge wishes to express. This use of social purposes to justify judicial decisions was introduced to AI and Law in [13] and more formally presented in [12]. Thus we may look to the law as one source for our value orderings: the assumption being that the moral order is at least compatible with the order reflected in legal decisions. Note that this legal order need not be static and may reflect changing social views and priorities. Although courts are supposed to be bound by precedents (the doctrine of *stare decisis*) as noted by Mr Justice Marshall in the US Supreme Court case of *Furman v Georgia* (408 U.S. 238 1972) there are occasions when "*stare decisis* would bow to changing values".

Several methods of deriving an audience, in the sense of a value ordering, from a set of cases have been proposed. In AGATHA [18] the value ordering which best explains a set of cases was discovered by forming a theory to explain a set of cases, and then attempting to provide a better theory, in terms of explaining more cases, until the best available theory was found. In [11], given a VAF and a set of arguments and a set of arguments to be accepted, the audiences (if any) to which that set is acceptable is determined by means of a dialogue game. Note that the ordering may not be fully determined (a *specific* audience): it may be possible that the desired set of arguments can be accepted by several audiences, represented as a partial order on the values. In [28], the VAF is rewritten as a meta-level argumentation framework [29], from which value orderings can emerge, or be formed, as a result of dialogue games based on the rewritten frameworks. In this last work explicit arguments for value orderings can be made in the manner of [30].

As well as legal cases, we can identify the approved value orderings from stories, using techniques for deriving character motives from choices with respect to actions, originally targetted at explaining the actions of people involved in legal cases [16]. Stories are often used to persuade people to adopt particular value orders, as with the fable and the parable we have considered in this paper. The notion of using didactic stories as arguments for value orderings was explored in [15] and [14]. Since stories like fables and parables were written specifically to advocate particular value orderings, they are highly suited to our purposes. The values concerned are typically clear, the choices sharp and the correct decisions clearly signposted, leaving little room for doubt as to the recommended preference.

We do not propose data mining or machine learning methods here. Although such methods can discover norms from a set of cases represented as facts and outcomes (e.g [32]), the discovered norms derive their authority from the amount of support in the dataset. They are suited to finding rules, but not exceptions, and it is exceptional cases, where norms need to be violated, that interest us. In law, however, single cases may form an important precedents, identifying apparent exceptions to existing norms, closing gaps and resolving conflicts,

[7] This would, of course, also justify the grasshopper stealing from the ant.

often revealing or choosing between value orderings as they do so.

As noted above, these methods may produce not a specific audience, but a set of audiences all of which conform to and explain the prevailing decisions. If this is so the question arises as to whether it is desirable or undesirable for all agents to be drawn from the same audience. To unify the audience would be to impose the designer's view as to what is moral, albeit constrained by the social decisions. In practice a degree of diversity may prove useful, leading to different agents occupying different social roles.

5 Summary

In this short position paper we have taken as our starting point the idea that as the use of agents speads and as they adopt the autonomous performance of ever more critical tasks, including perhaps, in the not very distant future, warfare and counter terrorism, there is a need to provide them with the capacity for moral reasoning. We have argued that neither of the approaches popular in current multi-agent systems, the enforcement of norms by the removal of the capability of violation, or the provision of a set of guiding principles will enable this. Moral behaviour requires and includes the recognition that on occasion it is right to violate norms, because while norms may be best observed in an ideal world, we need to be able to cope with the sub-ideal, and with the unforeseen. Unforeseen events may occur which mean that following a norm results in underdesirable effects, perhaps even subverting the very values the norm was designed to promote. Moreover when another agent transgresses norms, so producing a sub-ideal situation, it may be necessary to deviate oneself, either to punish the transgression or because the case is altered, and in the particular circumstances two wrongs *do* make a right.

But violation of a norm for moral reasons presupposes that the agent can recognise when the norm should be violated and what form the violation should take. This in turn requires that the agent be able to reason morally from first principles, by which we mean apply an ordering on values to the current situation. If we provide agents with a suitable value ordering, and the capacity to apply this value ordering when selecting an action, we can rely on the agents to make moral choices which might not be the case if they were to blindly follow a fixed set of norms. We have identified work which provides the basis for such a capacity. In doing so we provide a morality in the virtue ethics tradition of Aristole [3], as opposed to the consequentialism and deontology represented by current MAS approaches.

The literature also offers a number of approaches in which the moral orders for various societies can be derived from the legal decisions taken and the stories told in those societies. Note that we would expect both inter and intra cultural variation, and evolution over time.

Such matters can be explored and evaluated through simulations of the sort found in [26] and [27]. For a finer grained, qualitative evaluation, the techniques developed can be applied to classic moral dilemmas such as whether a diabetic may be allowed to steal insulin from another (the Hal and Carla case discussed in [19]) and Phillipa Foot's famous *Trolley Problem* [21].

Future work will need to investigate several aspects of value based reasoning, including: inducing value orderings; consideration of the extent to which values are promoted/demoted; and how value orderings can be applied to situations that differ (in some tangible way that suggests novelty) from the ones that originally gave rise to them.

REFERENCES

[1] Aesop, *Fables, retold by Joseph Jacobs*, volume Vol. XVII, Part 1, The Harvard Classics. New York: P.F. Collier and Son, 1909-14.

[2] Thomas Aquinas, *Summa theologica*, Authentic Media Inc, 2012, written 1265-74.

[3] Aristotle, *The Nicomachean Ethics of Aristotle, translated by W.D. Ross*, Heinemann, 1962, written 350BC.

[4] I. Asimov, *I, Robot*, Robot series, Bantam Books, 1950.

[5] K. Atkinson and T. Bench-Capon, 'Taking the long view: Looking ahead in practical reasoning', in *Proceedings of COMMA 2014*, pp. 109–120.

[6] K. Atkinson and T. Bench-Capon, 'Practical reasoning as presumptive argumentation using action based alternating transition systems', *Artificial Intelligence*, **171**(10), 855–874, (2007).

[7] K. Atkinson and T. Bench-Capon, 'Addressing moral problems through practical reasoning', *Journal of Applied Logic*, **6**(2), 135–151, (2008).

[8] K. Atkinson, T. Bench-Capon, and P. McBurney, 'Computational representation of practical argument', *Synthese*, **152**(2), 157–206, (2006).

[9] T. Bench-Capon, 'Persuasion in practical argument using value-based argumentation frameworks', *Journal of Logic and Computation*, **13**(3), 429–448, (2003).

[10] T. Bench-Capon, 'Transition systems for designing and reasoning about norms', *AI and Law*, **23**(4), 345–366, (2015).

[11] T. Bench-Capon, S. Doutre, and P. Dunne, 'Audiences in argumentation frameworks', *Artificial Intelligence*, **171**(1), 42–71, (2007).

[12] T. Bench-Capon and G. Sartor, 'A model of legal reasoning with cases incorporating theories and values', *Artificial Intelligence*, **150**(1), 97–143, (2003).

[13] D. Berman and C. Hafner, 'Representing teleological structure in case-based legal reasoning: the missing link', in *Proceedings of the 4th ICAIL*, pp. 50–59. ACM, (1993).

[14] F. Bex, K. Atkinson, and T. Bench-Capon, 'Arguments as a new perspective on character motive in stories', *Literary and Linguistic Computing*, **29**(4), 467–487, (2014).

[15] F. Bex and T. Bench-Capon, 'Understanding narratives with argumentation', in *Proceedings of COMMA 2014*, pp. 11–18, (2014).

[16] F. Bex, T. Bench-Capon, and K. Atkinson, 'Did he jump or was he pushed?', *AI and Law*, **17**(2), 79–99, (2009).

[17] D. Bošnački and D. Dams, 'Discrete-time promela and spin', in *Formal Techniques in Real-Time and Fault-Tolerant Systems*, pp. 307–310. Springer, (1998).

[18] A. Chorley and T. Bench-Capon, 'An empirical investigation of reasoning with legal cases through theory construction and application', *AI and Law*, **13**(3-4), 323–371, (2005).

[19] G. Christie, *The notion of an ideal audience in legal argument*, volume 45, Springer Science & Business Media, 2012.

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

[21] P. Foot, *Virtues and vices and other essays in moral philosophy*, Cambridge Univ Press, 2002.

[22] R. Hare, *Freedom and reason*, Oxford Paperbacks, 1965.

[23] H. Hart, *The concept of law*, OUP Oxford, 2012.

[24] A. Jones and M. Sergot, 'Deontic logic in the representation of law: Towards a methodology', *AI and Law*, **1**(1), 45–64, (1992).

[25] R. Lipsey and K. Lancaster, 'The general theory of second best', *The review of economic studies*, **24**(1), 11–32, (1956).

[26] M. Lloyd-Kelly, K. Atkinson, and T. Bench-Capon, 'Emotion as an enabler of co-operation.', in *ICAART (2)*, pp. 164–169, (2012).

[27] S. Mahmoud, N. Griffiths, J. Keppens, A. Taweel, and M. Bench-Capon, T.and Luck, 'Establishing norms with metanorms in distributed computational systems', *AI and Law*, **23**(4), 367–407, (2015).

[28] S. Modgil and T. Bench-Capon, 'Integrating object and meta-level value based argumentation', in *Proceedings of COMMA 2008*, pp. 240–251, (2008).

[29] S. Modgil and T. Bench-Capon, 'Metalevel argumentation', *Journal of Logic and Computation*, 959–1003, (2010).

[30] Sanjay Modgil, 'Reasoning about preferences in argumentation frameworks', *Artificial Intelligence*, **173**(9), 901–934, (2009).

[31] Ch. Perelman, *The new rhetoric*, Springer, 1971.

[32] M. Wardeh, T. Bench-Capon, and F. Coenen, 'Padua: a protocol for argumentation dialogue using association rules', *AI and Law*, **17**(3), 183–215, (2009).

[33] M. Wooldridge and W. van der Hoek, 'On obligations and normative ability: Towards a logical analysis of the social contract', *J. Applied Logic*, **3**(3-4), 396–420, (2005).