Artificial Intelligence manuscript No.
(will be inserted by the editor)

Norms and Value Based Reasoning: Justifying Compliance and Violation

Trevor Bench-Capon · **Sanjay Modgil**

February 11, 2017

Abstract There is an increasing need for norms to be embedded in technology as the widespread deployment of applications such as autonomous driving, 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. In this paper[1] we argue for the position that agents should be equipped with the ability to reason about a system's norms, by reasoning about the social and moral *values* that norms are designed to serve; that is, perform the sort of moral reasoning we expect of humans. In particular we highlight the need for such reasoning when circumstances are such that the rules should arguably be broken, so that the reasoning can guide agents in deciding whether to comply with the norms and, if violation is desirable, how best to violate them. One approach to enabling this is to make use of an argumentation scheme based on values and designed for practical reasoning: arguments for and against actions are generated using this scheme and agents choose between actions based on their preferences over these values. Moral reasoning then requires that agents have an acceptable set of values and an acceptable ordering on their values. We first discuss how this approach can be used to think about and justify norms in general, and then discuss how this reasoning can be used to think about when norms should be violated, and the form this violation should take. We illustrate how value based reasoning can be used to decide when and how to violate a norm using a road traffic example. We also briefly consider what makes an ordering on values acceptable, and how such an ordering might be determined.

1 Introduction

Norms are a topic of considerable interest in agent systems (Walker and Wooldridge (1995), Shoham and Tennenholtz (1997), Wooldridge and van der Hoek (2005), Moor (2006), van der

Trevor Bench-Capon
Department of Computer Science. University of Liverpool. E-mail: katie@liverpool.ac.uk

Sanjay Modgil
Department of Informatics. King's College, London.

[1] This paper is a revised, extended and consolidated version of several previous papers: Bench-Capon (2016a), Bench-Capon (2016b), Bench-Capon and Modgil (2016a) and Bench-Capon and Modgil (2016b).

Hoek et al. (2007), Sen and Airiau (2007), Savarimuthu et al. (2008), Ågotnes and Wooldridge (2010), Sugawara (2011), Mahmoud et al. (2015)). An early position paper, Moor (2006), set out to argue for the importance (and difficulty) of making machines behave ethically. Moor distinguished between implicit ethical systems (now usually termed *regimented* systems) and explicit ethical systems, based on rules and principles, and often employing deontic logics. This distinction still applies to most approaches to ethical or normative agents systems today. The importance of ethical systems, he argued, would grow as "future machines will likely have increased control and autonomy .... More powerful machines need more powerful machine ethics". Indeed this is now the case, as in the decade since Moor was writing, autonomous systems acting in the real world have become more and more a part of our reality, and recent successes in artificially intelligent technologies have prompted prominent researchers to argue for the urgency of research into ethical systems (Russell et al. (2016)).

Moor suggested that implicit ethical systems could be regarded as moral agents, even though "Of course, such machine virtues are task specific and rather limited. Computers don't have the practical wisdom that Aristotle thought we use when applying our virtues." But not all philosophers would agree that implicit ethical agents were ethical at all: Kant, for example, argues in Kant (1785) that actions done *in accordance with* the law only have moral worth if they are also done *for the sake* of the law. This seems correct: acting in accordance with what is right deserves no praise if it is impossible to act otherwise. None the less, perhaps, if we are unleashing agents in the real world, it is enough that they act in conformity with the law: moral worth is perhaps not an issue with machines.

However, regimented (Moor's 'implicit ethical') agents are designed to act in a limited, predictable environment, whereas the real world is not predictable, and unforeseen situations will arise. In these open systems, regimentation of behaviour by restricting agents to performing only permissible actions (e.g. Esteva et al. (2002), van der Hoek et al. (2007), Ågotnes et al. (2009)) does not allow agents to adapt to such unforeseen circumstances. Moreover, unlike norms found in legal and moral systems, such norms cannot be violated, and so it can be argued (e.g. Jones and Sergot (1992), Governatori (2015)) that they should not be seen as norms all, because the agents have no choice beyond compliance or non-participation. Such rules are thus more like the rules of a game, than moral and legal norms.

We therefore argue that agents in open systems should have the capacity to explicitly reason about the actions they perform, and in particular whether or not behaving according to normative prescriptions is in accordance with ethical criteria that these norms are designed to serve, so that if necessary agents may *choose* to violate norms. Again, as argued by Moor: "more powerful machines need more powerful machine ethics". However, this raises the question as to what are the ethical criteria by which 'explicit ethical agents' should choose which norms to adhere to. While quantitative utilitarian calculations may suffice in simple scenarios[2], more complex scenarios characteristic of real world moral dilemmas are not typically resolvable through such calculations (Dennet (1995)). What is required is a computational account of practical reasoning that equates with what Moor called the "practical wisdom of the type that Aristotle thought we use when applying our virtues". Practical reasoning has been investigated both within the standard multi-agent system context of beliefs, desires and intentions (BDI, e.g. Rao and Georgeff (1991)), in work such as Broersen et al. (2001), and, with ethical considerations at the forefront, through the combination of deontic, epistemic and action logics (van Den Hoven and Lokhorst (2002)). These approaches do not

[2] See for example a variation of the trolley problem setup (J. F. Bonnefon (2016)) in which a utilitarian autonomous vehicle will choose to direct itself so as to minimise the number of casualties.

directly address the issue of how to rationally choose amongst options for action by appealing to ethical criteria. However, since 2006 when Moor was writing, there has been growing interest in the use of argumentation theory in the context of practical reasoning in order to resolve such conflicts (e.g., Rahwan and Amgoud (2006)), and in particular Value-Based Reasoning and Argumenation as in Atkinson and Bench-Capon (2007) and Verheij (2016), in which the ethical criteria appealed to take the form of values promoted by actions and the relative importance (ranking) of these values. We will argue for the position that Value-Based Reasoning, especially in the manner of work based on the approach of Atkinson and Bench-Capon (2007), is one way to provide current and coming agents whose autonomy and operation will place them in unforeseen situations. where they must recognise the need to violate their norms, and to violate them in the most appropriate way, with the capacity to make such choices[3].

Of course, we do not advocate that value based reasoning and argumentation suffices on its own as a formalism for explicit ethical agents; comprehensive formal models may well require the integration of such reasoning with BDI and deontic logics, and possibly also the integration of quantitative and qualitative reasoning. Rather, the purpose of this paper is to propose that value based reasoning and argumentation is able to play an important role in equipping agents with the ability to reason about norms and related ethical considerations, and to therefore also be able to deal with unforeseen situations. We will support our position by illustrating use of value based formalisms for reasoning about norms using associated values characteristic of societies of increasing complexity, and how agents may then need to justify violation of norms in order to ensure promotion of the values that these norms were intended to serve.

We now further elaborate on some of the approaches mentioned above, by way of then introducing value based reasoning and argumentation. We will consider how value based reasoning can be used to justify norms given certain scenarios, and how differences in the values, preferences and state descriptions affect the norms that can be justified.

An excellent starting point for considering reasoning about norms is Ullmann-Margalit (1977), which does, of course, considerably pre-date multi agent systems, but none the less contains many relevant considerations. In that work, Ullmann-Margalit uses as scenarios on which to base her discussion of norms, simple two player games, such as the prisoner's dilemma (PD, Rapoport and Chammah (1965)). In such games there are two players and each can cooperate or defect, and their choices determine the payoffs. In PD as used in Ullmann-Margalit (1977), mutual cooperation gives a payoff of three to each player and mutual defection one to each player, while if the actions differ the defector receives five and the cooperator receives zero. Some key results concerning PD are that the Nash Equilibrium (Roth and Murnighan (1978)) is where both defect (since defection is the *dominant* action, and will receive the better payoff whatever the other player does) and that a successful strategy in iterated PD (where the players play one another repeatedly) is *Tit-For-Tat* (Axelrod (1987), but see Binmore (1998)). Using *Tit-For-Tat* an agent will cooperate in the first round,

[3] One of the anonymous reviewers stated "formal works in this field often aim to provide mathematical theorems, e.g., to relate the expressive power of a logic with the question of computational complexity that arise from this logic. … It is by no means obvious that mathematically oriented researchers in the field are committed to the claim that values should be left out. The fact that the focus has been on norms (in the crude sense of forbidding transitions) might have more to do with the present state of our formal modelling capabilities." Such formal concerns, and the need to overcome some of the limitations, are the basis of the REINS project described in Broersen (2014). We do not challenge this, but argue that the move towards Moor's full ethical agents, for example by the inclusion of the ability to reason to with values, and to enable them to violate norms, will be needed if autonomous agents such as driverless cars are to be deployed in practice.

and then copy its opponent's previous move in every subsequent round. Importantly PD is a non-zero sum game: the aggregate utility of mutual cooperation is greater than any other payoff, and the equilibrium in fact yields the lowest collective utility. Thus, it would in fact be mutually beneficial if one offered a payment to the other if they cooperated: this could secure a payoff of three and two, so that both would gain over mutual defection. Such agreements are, however, not possible in normal versions of the game, which do not allow for prior negotiations.

Public goods games have formed the basis of several studies of the emergence of norms in multi-agent systems such as Shoham and Tennenholtz (1997), Sen and Airiau (2007), Skyrms (2014), Bicchieri (2005), Sugawara (2011) and Mahmoud et al. (2015). Empirical studies suggest, however, that applying standard game theory[4] to public goods games does not provide a very realistic model of actual human behaviour. Experiments using such games are very common and have formed the subject of a number of metastudies. For example Engel (2011) examined 131 examples of the Dictator Game and Oosterbeek et al. (2004) was based on 37 papers reporting Ultimatum Game experiments. In none of these many studies was the canonical model of game theory and economics[5] followed. Although the study of Henrich et al. (2001) looked at fewer societies (fifteen), it is particularly interesting in that the studies considered highly homogeneous societies, and used the same methodology for each experiment. Again none of the societies followed the canonical model. Consequently it is not plausible to see the canonical game theoretic models, and their criteria such as the Nash Equilibrium, as justification for the norms encountered in such societies.

Another approach is to model scenarios as State Transition Diagrams (STD), and to investigate how norms can be designed in such situations to avoid unwanted states of affairs as in Wooldridge and van der Hoek (2005) and Ågotnes and Wooldridge (2010). In these approaches, agents are typically represented using the Belief-Desire-Intention (BDI) model (Rao and Georgeff (1991), Wooldridge (2009)), inspired by Bratman (1999). The BDI model supposes agents to have a set of *beliefs* and a set of dispositional goals (*desires*). Actions are chosen by identifying the desires that can be realised in the current situation (candidate *intentions*), and then committing to one or more of these intentions, and choosing a course of action intended to realise the associated goals. This, however, leaves open the question of where the desires come from in the first place. This in turn means that in BDI systems, there is no explanation of where goals come from. Often they are completely fixed, and even in systems where they can be derived from the current state (e.g. Rahwan and Amgoud (2006)), there is a fixed set of potential desires some of which are active in a given situation, which means that we can say why the desire is active, but not why there is such a desire to activate. This inability to justify desires greatly limits the ability to engage in moral reasoning in a transparent fashion.

This paper proposes use of an alternative approach to action selection, often called *practical reasoning* (Raz (1979)) and formalised as Value-Based Reasoning, as a means to enable agents to justify norms by reasoning about the social and moral *values* that norms are designed to serve; that is, the kinds of moral reasoning we expect of humans. Agents are associated with a set of social values, the aspirations or the purposes an agent might pursue, such as liberty, equality, fraternity, wealth, health and happiness. These values provide reasons why certain situations are considered goals by the agent, and so allows for the justification of "desires", as explained in Atkinson and Bench-Capon (2016). The basic idea is

[4] In which, as in classical economics, players are all rational, self-interested and perfectly informed and act so as to maximise their own utility.

[5] In which, in the example of the Ultimatuatum Game, the proposer would offer the smallest amount possible, and the recipient would accept any offer, no matter how small.

that agents have a set of such values and their aspirations and preferences are characterised by their ordering of these social values. Attempting to promote (or avoid the demotion of) preferred values provides reasons for desiring some states of affairs, and for choosing particular actions. In the context of argumentation based models of reasoning, an ordering on values can be conceptualised as an *audience* to which the arguments are addressed (Perelman (1971)), and acceptance of an argument as to what to do depends not only on the argument itself - for it must, of course, be a sound argument - but also on the audience addressed. This notion of audience as an ordering on values was computationally modelled in Grasso et al. (2000) and made more formal in Value-Based Argumentation Frameworks (VAFs, Bench-Capon (2003)). VAFs are an extension of the abstract Argumentation Frameworks (AFs) introduced in Dung (1995), but whereas in an AF an argument is defeated by any attacking argument, in a VAF an argument is *defeated for an audience* by an attacker only if the value associated with the attacking argument is ranked at least as highly as the attacked argument by that audience. In this way different audiences will accept different sets of arguments (preferred semantics (Dung (1995)) is used to determine acceptance), and, as is shown in Bench-Capon (2003), provided the VAF contains no cycles in the same value, there will be a unique non-empty preferred extension.

Use of VAFs provides a way of explaining (and computing) the different arguments accepted by different audiences. Value-Based Reasoning has been used as the basis of practical reasoning in, amongst others, Garcez et al. (2005), Atkinson and Bench-Capon (2007), and van der Weide et al. (2011), and applied in particular areas including law (Bench-Capon et al. (2005)), e-democracy (Cartwright and Atkinson (2009)), policy analysis (Tremblay (2016)), medicine (Atkinson et al. (2006)), experimental economics (Bench-Capon et al. (2012)), rule compliance (Burgemeestre et al. (2011)), decision support (Nawwab et al. (2008)) and even ontology alignment (Trojahn et al. (2008), Payne and Tamma (2015)). Complexity results for VAFs were established in Dunne (2010) and Nofal et al. (2014). Here we will discuss norms and their design and justification in terms of the Value-Based approach to practical reasoning.

In Section 2 we review background work on the formalisation of Value Based Argumentation in Alternate Action Based Transition Systems (AATS) (Wooldridge and van der Hoek (2005)). These systems model open agent systems (qua models of 'worlds' in which agents engage in joint actions to bring about desired states of affairs). Section 3 then shows how this approach enables reasoning about and justifying the norms that serve values and value orderings characteristic of societies of increasing complexity. Section 4 further explores the use of such reasoning by agents who may need to justify violation of norms in order to ensure promotion of the values that these norms were intended to serve. We also briefly consider what makes an ordering on values acceptable, and how such an ordering might be determined, in Section 5. We conclude with some reflective discussion and point to future work in Sections 6 and 7.

2 Background

In this section we provide some essential background: the structure which we use to model our "world", Alternate Action Based Transition Systems (AATS); the value-based arguments that agents can generate from such structures and use to justify their actions in this environment, and; the running example we will use to illustrate our model.

2.1 Alternate Action Based Transition Systems

In open agent systems, an individual's choice does not necessarily determine the state that will be reached. To account for this, open agent systems should model transitions as the *joint actions*[6] composed of the individual actions of all the agents relevant to the situation. A suitable variant of state transition diagrams for use in open agent systems is *Action-based Alternating Transition Systems* (AATS), introduced in Wooldridge and van der Hoek (2005), since their transitions are the joint actions of all the agents relevant to the situation. AATS are formally based on Alternating-time Temporal Logic (Alur et al. (2002)). The basic AATS was augmented in Atkinson and Bench-Capon (2007) to allow the labelling of the transitions with the values promoted and demoted by that transition. AATSs labelled in this way were termed *Action-based Alternating Transition Systems with Values* (AATS+V) and AATS+Vs were used to provide the underpinning semantical structure for the approach to practical reasoning set out in that paper. The formal definitions are given in the following subsection.

2.1.1 Formal Definitions

Definition 1: AATS (Wooldridge and van der Hoek (2005)). An *Action-based Alternating Transition System* (AATS) is an $(n + 7)$-tuple $S = \langle Q, q_0, Ag, Ac_1, \ldots, Ac_n, \rho, \tau, \Phi, \pi \rangle$, where:

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

AATSs are particularly concerned with the joint actions of the set of agents Ag. j_{Ag} is the joint action of the set of n agents that make up Ag, and is a tuple $\langle \alpha_1,...,\alpha_n \rangle$, where for each α_j (where $j \leq n$) there is some $ag_i \in Ag$ such that $\alpha_j \in Ac_i$. Moreover, there are no two different actions α_j and $\alpha_{j'}$ in j_{Ag} that belong to the same Ac_i. The set of all joint actions for the set of agents Ag is denoted by J_{Ag}, so $J_{Ag} = \prod_{i \in Ag} Ac_i$. Given an element j of J_{Ag} and an agent $ag_i \in Ag$, ag_i's action in j is denoted by j^i. This definition was extended in Atkinson and Bench-Capon (2007) to allow the transitions to be labelled with the values they promote.

Definition 2: AATS+V (Atkinson and Bench-Capon (2007)). Given an AATS, an AATS+V is defined by adding two additional elements as follows:

[6] Here, as in Alur et al. (2002) and Wooldridge and van der Hoek (2005), by *joint action* no implication of the agents acting in concert is intended. A joint action is simply an action composed of actions performed by a set of agents at the same time, without any suggestion of coordination, or common purpose. This contrasts with the notion of joint action in e.g. Levesque et al. (1990), which concerns acting in teams.

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

An *Action-based Alternating Transition System with Values* (AATS+V) is thus defined as a $(n + 9)$ tuple $S = \langle Q, q_0, Ag, Ac_1, ..., Ac_n, \rho, \tau, \Phi, \pi, V, \delta\rangle$. The value may be ascribed on the basis of the source and target states, or in virtue of an action in the joint action, where that action has intrinsic value[7].

2.2 Reasons for Action

The values give agents reasons to perform or not to perform the various actions, based on the argumentation scheme proposed in Atkinson and Bench-Capon (2007). A number of such reasons are given in Atkinson and Bench-Capon (2014) (the "N" suffix denotes reasons not to perform the action: j is a joint action in which agent ag performs j^{ag}, ϕ is a *goal*, which holds or fails to hold in a given state, and which agents may attempt to realise, maintain, avoid or remove).

R1 We should participate in j in q in which ϕ holds to maintain ϕ and so promote v.
R2N We should not participate in j in q in which ϕ holds since it would remove ϕ and so demote v.
R3 We should participate in j in q in which $\neg\phi$ holds to achieve ϕ and so promote v.
R4N We should not participate in j in q in which $\neg\phi$ holds since it would avoid ϕ and so fail to promote v.
R5 We should participate in j in q to ensure ϕ and so promote v. Note that ϕ may be contingently realised or unrealised in q and that, in some variants, the promotion of v might not be immediate, or permanent. This also applies to R5N and R6.
R5N We should not participate in j in q which would ensure $\neg\phi$ and so demote v.
R6 We should participate in j in q to prevent $\neg\phi$ and so promote v. Note that $\neg\phi$ may be contingently realised or unrealised in q.
R6N We should not participate in j in q which would prevent ϕ and so fail to promote v. We suggest that to make the reason worth consideration we should only use variants which prevent ϕ immediately and permanently.
R7 We should participate in j in q in which $\neg\phi$ to enable ϕ to be achieved and v to be promoted on the next move.
R8N We should not participate in j in q in which ϕ which will risk ϕ being removed on the next move which would demote v.
R9 We should participate in j in q because performing j^{ag} promotes v.
R9N We should not participate in j in q because performing j^{ag} demotes v.

Objections to these arguments can be formed by questioning whether the state is as claimed, the consequences of the action will be as specified, whether the goal is realised and whether the value is indeed promoted. The arguments, attacks and rebuttals of the attacks are then organised in a Value-Based Argumentation framework (VAF) as described in Bench-Capon (2003) and evaluated according to an ordering on the values. These value orderings

[7] Although the labellings are taken as givens in Atkinson and Bench-Capon (2007), it is possible to make the ascription of values to transitions transparent, and to enable agents to argue about what values should label a given transition. This was fully discussed in Atkinson and Bench-Capon (2016).

will depend on the subjective preferences of the particular audience, and so different agents may quite rationally choose different actions.

To summarise: three stages in practical reasoning are identified in Atkinson and Bench-Capon (2007):

- **Problem formulation**: essentially the construction of an AATS+V for the particular problem situation. The AATS+V will reflect the views of the agent engaged in the reasoning, and so will embody that agent's causal model (to determine the transitions) and its values (to enable the labelling of transitions), as is demonstrated in Atkinson and Bench-Capon (2016). As Atkinson and Bench-Capon (2016) indicates there can be arguments justifying the formulation of the problem, before the AATS+V is used in the subsequent stages.
- **Epistemic stage**: this involves determination of what the agent engaged in the reasoning believes (or chooses to assume) about the current state and the joint action that will result from the choice of a particular individual action by the agent concerned;
- **Option selection**: the arguments generated from the AATS+V based on the reasons given above, and objections and counterexamples to these arguments, are formed into a VAF and their acceptability status determined according to the preferences of the agent engaged in the reasoning. The acceptability status determines which actions, according to its beliefs, values and preferences, the agent can justifiably choose to perform.

In the work mentioned above, and in the examples below, value preferences are considered pairwise. In fact transitions may be labelled with several values, so that the comparison should be between sets of values: our approach is just to use the most preferred of these values. Use of sets of values was considered in Prakken (2002), Chorley and Bench-Capon (2005) and Bench-Capon et al. (2011). Also it might be desirable to consider degrees of promotion of values, as advocated in Sartor (2010). Such extensions, however, still require further work if a consensus as to the best approach is to be achieved, and, since their use does not affect the general principle of providing agents with the capacity to perform value-based reasoning, we will continue with the simple pairwise comparisons in this paper.

2.3 Example

An AATS+V was used in Bex et al. (2014) to model the states and actions found in both the fable of *The Ant and the Grasshopper* (Aesop (1909)) and the parable of *The Prodigal Son* (Luke 15:11-32). Fables and parables are suitable examples for us because they are stories with a moral point, and are frequently used in the moral education of children. In *The Ant and the Grasshopper*, the story is that during the summer the grasshopper sings and plays while the ant works hard storing up food for the winter. When the winter comes, the grasshopper has no food; nor will the ant give away any of its store, and so the grasshopper dies. In *The Prodigal Son* the prodigal wastes his inheritance on idle play but when destitute asks his father for forgiveness; the father does forgive and takes him back into his household.

An AATS+V based on the model of Bex et al. (2014) is shown in Figure 1. In our example, food is sufficiently abundant in summer that one can gather food and eat without effort. Growing food for the winter is, however, a full time effort (digging, planting, weeding, reaping, storing) and produces a surplus, but the nature of the activity is that it is either done or not: the amount produced is not proportional to the effort. The food cannot be stored over the summer: therefore the winter ends with a period of carnival (q5, q8 and q12) when the surplus is consumed with feasting. The state has five propositions. The first two indicate

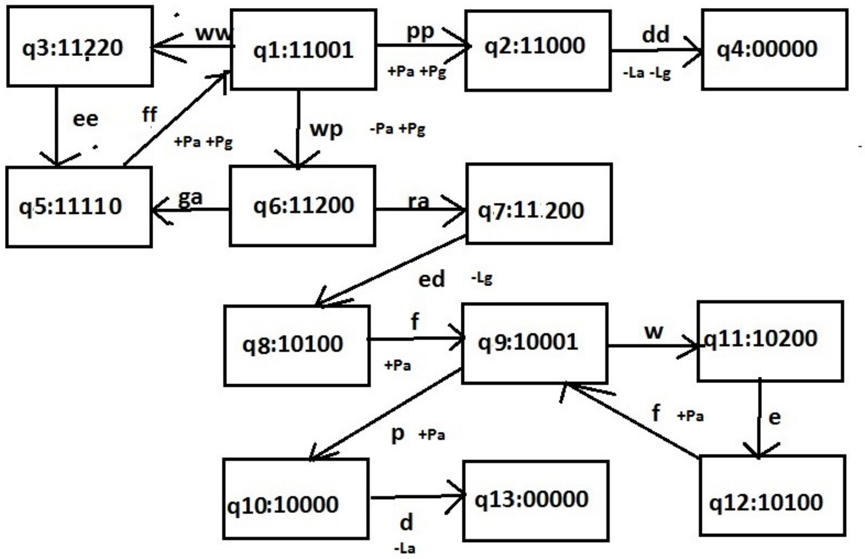

Fig. 1 AATS+V for the Example: w = work, p = play, a = ask, g = give, r =refuse, e = eat, f = feast d =die. The same AATS+V is used for both the fable and the parable. Joint actions give that action of the ant/father, followed by that of the grasshopper/son. States are vectors of five propositions representing: ant/father alive, grasshopper/son alive, ant/father has food. grasshopper/son has food, season is summer, respectively.

whether the ant (father) and the grasshopper (son) are alive, the third represents whether the ant (father) has no food (0), enough (1) or abundant food (2), the fourth similarly represents whether the grasshopper (son) has no, enough or abundant food, and the fifth whether it is summer (1) or winter (0). Thus the initial state, q_1 is 11001, indicating that both are alive, neither has food and that it is summer. Now, if both work and the state becomes q_3, 11220 indicates that both are alive, both have abundant food and that it is winter. The key decisions are in the initial state (q1) where the grasshopper and the prodigal choose to play (action *p*) while the ant and the father choose to work (action *w*), thus transitioning to q6, and in q6 where the ant refuses the grasshopper (action *r*), so transitioning to q7, while the father gives to the prodigal (action *g*), so transitioning to q5. In the other states there are no choices to be made.

We have labelled the diagram in Figure 1 with just four values. Life for the ant (father) and grasshopper (son) (L_a and L_g) and Pleasure for the ant (father) and the grasshopper (son) (P_a and P_g). Additional value labels will be identified as our discussion proceeds.

3 Justifying Norms with Value Based Argumentation

The basis of our approach is that agents will be faced with situations requiring a choice of action. They will normally choose the action which has the most favorable consequences for themselves; in our terms the action which promotes their most preferred value or set of values. Since different agents will have different preferences they will, quite reasonably, make different choices. But agents do not exist in isolation: their choices will impact on other agents, and what other agents choose will affect the consequences of their own actions. In a social group norms will develop, or be imposed. These are important where the choices of individuals may go against the interests of the group, or where additional information about how others may be expected to act is useful. Norms may be social conventions or laws. In both these cases they are typically reinforced by sanctions (see Bench-Capon (2015)). Legal norms are typically backed by formal sanctions such as fines and imprisonment, whereas social norms are backed by informal sanctions such as ostracism, ridicule or perhaps just disapproving looks. There are also moral norms. Conventions and norms may embody moral norms, and should at least be consistent with them, but unless so embodied, moral norms are not usually associated with sanctions. Conventions and laws, introducing as they do the notions of sanctions and punishments, will often be associated with a purpose, such as to promote the sustainability of the group or society. Simulations such as Mahmoud et al. (2015) have shown that punishments are essential if normative collapse is to be avoided: more than that, stability requires that failure to punish be itself punished.

In this section we consider how a set of norms might develop from a consideration of value-based practical reasoning in the example scenario.

3.1 Arguments in q_1

We first consider the arguments as to what to do, available to an agent in q_1, based on the values of pleasure and life. The agent's own pleasure and life will be denoted P_s and L_s, the pleasure and life of the other as P_o and L_o. Our arguments are derived from the reasons of section 2.2, but here expressed in terms of only the agent's own action and the value, e.g. *you should perform α since it will promote v*, where α is the agent's action in the justified joint action, and v is the value promoted.

A You should not play since it will risk L_s being demoted (R4N)
B You should work since it will enable P_s to be promoted (R7)
C You should play to promote P_s (R9)
D You should not work since it will demote P_s(R9N)

Thus we have reasons pro and con working: the pro reason is the future pleasure it enables (**B**), and the con reason is the immediate loss of pleasure which accompanies it (**D**). Play in contrast affords immediate pleasure (**C**), but risks the loss of life (**A**). The risk associated with argument A is substantial: avoiding death requires both that the other works, and that the other will forgo its own pleasure in order to save one's life. Therefore (assuming life is preferred to pleasure) only the most risk taking individuals will choose to play in q_1.

In the following subsection we will show how this norm can also be explained and justified in terms of value based reasoning.

3.2 A first set of norms

So let us suppose that the society has the norm:

SN1[8]: It is forbidden to play

[8] We use SN for social norms, which can be moral norms, but may also be found, backed with sanctions, as social and legal norms. We use MN for those norms better seen as purely moral norms, those typically not incorporated in a social or legal code, and not typically associated with sanctions.

If all agents conform to this norm the situation can continue indefinitely round the loop q_1,q_3, q_5, q_1. But there will always be the temptation to violate SN1: the value L_s is threatened, and if q_6 is reached, there is the possibility that the other agent will give the required food. Work on norms such as that reported in Axelrod (1986) and Mahmoud et al. (2015) suggests that unless there is some reinforcement mechanism, norms are liable to break down. The reinforcement mechanism is to require violations to be punished by other members of the group, which in q_6 would mean that food is withheld. Moreover, to avoid the normative collapse demonstrated in Mahmoud et al. (2015), it is necessary to punish those who do not punish, and so punishment needs to be seen as obligatory. This in turn means that we need a norm applicable in q_6:

SN2: It is forbidden to give food

Refusal to give food could be justified without reference to any norms by an argument based on R6N, *you should not give since that will fail to promote P_s*. Given the counterargument based on R5N, you should not refuse since this will demote L_o, and so refusal requires a preference for P_s over L_o. But this does seem selfish rather than moral, and acts against sustainability. It does not seem morally good to prefer a minor value in respect of oneself to an important value in respect of the other: in Atkinson and Bench-Capon (2008), for example, moral behaviour was characterised as not regarding lesser values enjoyed by oneself as preferable to more important values in respect of others, which would oblige the agent to give. We can however introduce another value, Justice (J), which will justify refusal. This has some intuitive appeal, since the foodless agent has chosen to be in that position, and is attempting to take advantage of the efforts of the other. Thus recognising justice as a third value (labelling the transition q_6-q_7), preferred to L_o, will justify the punishment of violations of SN1. This would be difficult to justify under a simple consequentialist perspective proposed in Bentham (1825) (since it means the grasshopper dies), although it might be possible to recognise Justice in more sophisticated versions of Utilitarianism by considering the effects of additional, more remote, consequences of normative collapse and those resulting from excessive free loading[9]. It is more straightforwardly incorporated in virtue ethics (Aristotle (1962)) or deontological (Kant (1785)) approaches since it is capable of universal adoption, and it is not difficult to see a preference for justice as virtuous, since it can be justified in terms of equity and sustainability, by preventing freeloading and the collapse of SN1.

The result will be a pair of norms which are capable of resisting collapse, according to the empirical findings of Mahmoud et al. (2015). The result is a rather puritan society (relieved only by a brief period of hedonism), based on subsistence farming, with a strong work ethic, and an aversion to what in the UK is currently termed a "something for nothing" society. An alternative would be to introduce a fourth value, Mercy, that labels the transition $q6$-q_5, and is preferred to Justice. This preference is very possibly the recommendation of the parable of *The Prodigal Son*, and would also allow society to continue, at the expense of a sacrifice of pleasure by the ant/father. But it is a feature of the parable that the son repents, and there is a tacit understanding that the son will not repeat the pattern, but will choose work in future. We might therefore wish to modify SN2 to something like:

- **SN2a** It is allowed to give food only once,

which would permit the giving of food. We might wish to go further and to accompany this with:

[9] Mill may himself have foreseen the problem and arguably advocated rule rather than act utilitarianism: see Urmson (1953).

- **SN2b** It is obligatory to meet the first request for food.

This would represent a preference for Mercy, but enforce *a two strikes and you are out* policy, so that Justice is still respected. It also opens the possibility for the ant to play at the grasshopper's expense on some future cycle (cf. children supporting elderly parents). Whereas simply removing SN2 would lead to the possibility of exploitation, and so devalue Justice, the pair of norms SN2a and SN2b retain due respect for Justice, while allowing scope for Mercy until the object of the mercy proves incorrigible. This will require the state vector to have an extra term to record violations.

We now consider developments beyond the basic scenario of Figure 1. These will necessarily be described in less detail, both because of space limitations and because of the large number of possible variations. Of course it would, in future, be desirable to extend the scenario at the same level of detail and provide an expanded AATS+V modelling a more complex society. This would be a very substantial undertaking: we hope that it is clear that the discussion in the following sections is making use of the aforementioned techniques.

3.3 Critique

Although the norms SN1 and SN2 (with or without the variants SN2a and SN2b) will give rise to an equitable and sustainable society, we might expect to see thinkers in such a society as questioning the worth of the society. There are a number of grounds for critiques. For example:

- There is no net pleasure in the society: the displeasure of working is offset by the pleasures of feasting at carnival, but there is no net gain. Such a society lacks progress and reward and any point beyond it own continuance.
- There is no choice or diversity in the society: the path taken is determined at all times by the norms, and deviation is effectively punishable by death.
- The pleasure enjoyed in this society is of a rather basic kind, whereas the pleasure it denies itself might be seen as a higher pleasure. The hard line utilitarian might adopt the view of Bentham (1825) that "Prejudice apart, the game of push-pin is of equal value with the arts and sciences of music and poetry", but others, like Mill would disagree: "it is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied" (Mill (1871)). Such higher pleasures can only be provided by (certain forms of) play, not by feasting.

Therefore critics might seek a way of changing the normative code so as to improve the society in one or more of these respects. Certainly, it is considered essential to a civilised society that it is able to generate a surplus of food and so allow scope for the development of arts and sciences, or even the simple enjoyment of leisure time. There is therefore some push behind finding a justification for allowing some relaxation of the rigid society represented by SN1 and SN2. In order to further the discussion we will distinguish between three types of pleasure, by introducing different values for different pleasurable activities. We will retain P for the bodily pleasures associated with carnival (feasting and the like): toil will also continue to demote this value. We will also distinguish between approved activities made possible by not working (e.g. arts and sciences) which we will term *culture* (C), and deprecated activities (e.g. gaming or mere idleness) which we will term *frivolity* (F). We thus need to distinguish between approved play ($play_a$, i.e engagement in culture producing activities) and deprecated play ($play_d$, i.e. engaging in frivolity). We might even modify SN2b to give food only to someone in need because of $play_a$, and to withhold food from those in need as a result of $play_d$.

3.4 Allowing For Play and the Evolution of More Complex Societies

There are a number of ways in which we can accommodate agents who play rather than work. Some require disparity between agents, while others require a redescription of the world, additional values and a finer grained description of activities and states.

3.4.1 Power

We first consider a disparity of power. In this situation some agents are sufficiently more powerful than the others to be able to compel them to surrender their food. We assume that the powerful agents comprise less than half the population. This is modelled by changing the *ask* action for the powerful agents allowing them to *demand*, rather than *request*, food in q_6, and to render it impossible to refuse a demand, so that there is no *ra* transition between q_6 and q_7. This removes argument A for the powerful, since there is no longer any risk in playing because their demands must be acceded to. Might the powerful play and demand all the food from the ant so that they can also feast? This would result in the ant starving and so would be a short term expedient, since the ant would die and the powerful be forced to work in subsequent years. So we should perhaps have a norm to prevent this:

SN3 It is forbidden to demand non-surplus food.

This can be based on a value preference for the Life of the other over Pleasure.

Remember now that we have distinguished between three types of pleasure so that argument C needs to be split into two arguments:

- **C1**: You should $play_a$ to promote culture (C).
- **C2** : You should $play_d$ to promote frivolity (F).

Now the powerful will not choose to work unless they prefer their future pleasure (P_s) to the current C and F afforded by play. They also have a choice of leisure activity, depending on whether they prefer culture or frivolity. Of course, a moral preference is built into the names of the values, and there will be a moral norm, *applicable only to the powerful*:

MN4 It is forbidden to $play_d$.

Whether MN4 is adopted as a social norm will depend on whether the peer group of the powerful see themselves as having also a social obligation not to indulge in idle pleasures (*noblesse oblige*[10]). MN4 allows the choice to work to be morally acceptable, since it is comfortable with a preference for physical pleasure over culture, forbidding only idle pleasures. Alternatively we can represent Mill's position that only the higher pleasures are morally worthy with MN4a:

MN4a It is obligatory to $play_a$

[10] It is very uncommon for a norm like MN4 to be enacted as law. The sanction of social ostracism is typically seen as sufficient deterrent to keep the frivolous down to a sufficiently low number.

(also directed only at the powerful). The problem with both these cases is that this means that there is one norm for the powerful and one norm for the powerless. To justify this distinction, there needs to be some kind of social order, recognised by all, and considered morally acceptable, so that the class difference between those able to demand food in q_6 and those not so able is seen as acceptable. This is not at all unusual in actual societies: for example Mrs Alexander's well known hymn *All Things Bright and Beautiful*, often seen as particularly directed towards children, contains the verse (seldom sung these days):

> "The rich man in his castle, The poor man at his gate, God made them high and lowly And ordered their estate."

This equates with a preference ordering $L \succ C \succ P \succ F$ that can be advocated within a Virtue Ethics approach or a Consequentialist approach (indeed given Mill's view that not all pleasures are of equal worth, the consequences are an improvement: since only the powerful *can* act to promote culture, it is good that they do so, even if it is at the expense of the powerless, since culture is preferred to pleasure).

One example of such a society is Feudalism. A good model for such a society is where some agents own the land and allow tenant farmers to work the land in exchange for the payment of rent. The nature of such a society is coloured by the ratio of powerful to powerless. If there are relatively few powerful, they can demand low rents and so leave some surplus to the tenants and allowing some degree of feasting to them (so shortening rather than removing the carnival period). This will also mean that there will be some surplus food available after the needs of the powerful have been met, some of which can be demanded to give the powerful pleasure as well as culture.

What is important, for the sustainability of such a society, is that the powerless respect the social order and do not rise up and overthrow the elite. Revolutions must be avoided. The social order can be reinforced by including a value *deference* (D), promoted by working if one has no power to demand, and by giving when food is demanded, and so promoted by the transitions q_1-q_3 and q_6-q_5. This gives the powerless arguments to respect the social order, to "know their place". Deference can reinforce the preference for C over F by being seen as promoted by the transition using $play_a$ and *work*, but not the transition $play_d$ and *work* (the idle masters do not command respect). This value recognises two different roles: the powerful are required to promote culture (MN4a) and the rest are required to enable them to do so. Acceptance can be reinforced in several ways including patriotism, in which the powerless are encouraged to take pride in the cultural achievements of their masters; or religion, as in the hymn quoted above. As a further reinforcement, prudence suggests that the rents should not be too high, since that is likely to push the workers in a revolutionary direction.

A further possibility is that some workers may be taken out of food production and used for other purposes of benefit to all, which might be additional cultural activities (e.g. minstrels), building works (e.g. the pyramids), or whatever, and then fed from the tribute. Thus, once the despot's own needs have been considered, the surplus can be apportioned by them between allowing some retention by its producers and some public works ("bread and circuses"). In the models, the fewer in the powerful class the greater the scope for ameliorating the lot of the powerless, and hence the more likely the society is to be stable. This also appears to be supported by some historical examples. In feudal societies it seems that the powerless suffer more when there is a weak king and squabbling barons than when there is a powerful king who keeps the barons in check[11]. The proportion that is taken has

[11] For example, in the Robin Hood legends the people favour Richard over his weak brother John.

been investigated in behavioural economics Loewenstein (1999)[12]. At the limit, where the classes are equally divided, there is no leeway: there the powerful require all the surplus.

In addition to Feudalism, there are other models: slavery is one, and the kind of brigandry depicted in the film the *Magnificent Seven* is another. But these afford far less opportunity for keeping the powerless content, and so are liable to breakdown. In the film the banditry is stopped by force, and slavery was abolished, whereas Feudalism evolved into a different social order, rather than being abolished or overthrown (at least in the UK: in France things were ordered differently). The key distinction is that the powerful restrain their demands so that revolution is not seen as worthwhile[13]. To reinforce this, we often find notions of "*noblesse oblige*", which, as remarked above, attempts to persuade the powerful in the direction of admirable forms of recreation and philanthropy. We will term the associated value as *generosity* (G), and it is the *quid pro quo* of deference. This might form the basis of the moral norm:

MN5 It is obligatory to be generous in your treatment of the less fortunate

and the ordering: $L \succ C \succ G \succ D \succ J \succ P \succ F$. We still need $C \succ G$ because the point of this social order is to permit $play_a$. G is there to encourage stability, not as an end in itself. Note that, part of this accommodation is to play down which persons actually enjoy the various pleasures. Culture is now seen as a public good and $play_a$ a duty. People are expected to promote the values they can, given their social position. We have accordingly omitted the suffices indicating beneficiaries. Note that MN5 is rarely a legal norm when there is an imbalance of power: it will be at most adopted as a peer group convention. This may be because of the difficulty of enforcing sanctions on powerful transgressors, short of revolution. Note, however, that Generosity could lead the powerless to give away food to the needy: it could replace Mercy as a motivation for SN2a and SN2b.

3.4.2 Wealth

In post-feudal societies we find that class and disparity remain, but that this disparity is manifested as wealth rather than physical coercion. In a sense this transition began in the feudal age, when power began to take the form of (enforceable) land ownership rather than force of arms.

When wealth is the source of power, the forcibly coercive demands of the powerful are replaced by the ability to buy the surplus. So here the transition between q_6 and q_5 becomes *buy* and *sell* rather than *ask* (or *demand*) and *give*. In this model, selling is not compulsory and so the possibility of reaching q_7 is there. However not selling restricts the hoarder to promoting P and jeopardises L_o, whereas selling not only avoids demoting L_o, but also opens up the possibility of enjoying some $play_a$ or even $play_d$. For example, by selling half the surplus for two cycles, a worker would be able to save so as to accumulate sufficient wealth to spend the third in play of one or the other kinds and then buy food for the winter. This is the underlying idea of holidays, pensions, and more recently of "gap years". The

[12] The powerful find themselves in the position of the Dictator in the Dictator Game, or Proposer in the Ultimatum Game. Both of these have been much studied in behavioral economics (Engel (2011) and Oosterbeek et al. (2004)). These studies have suggested that it is rare for people to keep as much as they can for themselves, and that Respondents in the Ultimatum game will take nothing if offered less than what they consider to be a fair amount. Explanations for behaviour in the two games in terms of value-based argumentation can be found in Bench-Capon et al. (2012).

[13] In the words of the blues song *Custard Pie Blues* by Sonny Terry and Brownie McGhee "You have to give me some of it, or I'll take it all away".

balance between how the surplus is distributed between *work*, $play_a$ and $play_d$ can be left to the individuals and so made to depend on the preferences of individuals, or there may be norms imposing limits. At this point it is useful to distinguish between values that are maximisers, for which more is always better, and values which are satisficers[14] for which enough can be enough and more is of no benefit and possibly of harm: for example, one will become sated without too much feasting[15].

In its simplest form, this model should lead to a fair degree of equality, since eventually the initially wealthy will have spent all their money, and so be forced to work, since there is no other source of income. There are, however, mechanisms which tend to allow the wealthy to maintain their position:

- The wealthy may own the land (or the means of production) and be in a position to take some proportion of the labour of others in the form of rent or profit. The situation is little different from the feudal, except that payment is now in money, not in kind. The flexibility afforded by money is more suitable to an Industrial society where production requires more than land and labour, and where produce is not bread alone, but a whole range of manufactured goods.
- The wealthy may choose to lend money at interest. Since many will regard a "bird in the hand as worth two in the bush", there is likely to be takers for such loans, allowing for people with initial wealth to pay for their needs from the interest and maintain their wealth, and perhaps even, given sufficient borrowers or high enough interest rates, to increase it. Note, however, this requires some way of ensuring that the lenders can be confident that the interest will be paid, and the debt repaid. This in turn requires some kind of norm, e.g.

  SN6a It is obligatory to repay debts.

  This would be associated with a new value of *Trustworthiness* or *Honesty* (H), promoted by observance of debts (and contracts and agreements generally) and demoted by reneging on such agreements. Such agreements are typically governed not by moral norms alone, but are supported by laws and enforced by sanctions. In order to make this more general we might prefer to use the formulation:

  SN6 It is obligatory to honour agreements.
- Some people may have access to wealth from outside. For example, in the sixteenth century the Spanish rulers had a seemingly inexhaustible supply of gold and silver from the Americas.
- Deference or Generosity may mean that some agents are not required to work or pay but are simply given some kind of tribute. For example monks or priests may be supported by tithes or donations, or the infirm by alms. The latter, where the motivating value is generosity, are perhaps covered by MN5, but modified to be applicable to all, and to cover all those considered worthy of alms, not just the unfortunate. Moreover, the notions of tithes (and today contributions to welfare benefits) are often the subject of law, rather than left to morality.

  SN5a It is obligatory to give alms to those unable, for good reason, to support themselves.

  This rephrasing as SN5a means that we broaden the notion of *unable to support themselves* from incapacity to include those engaged in some other, worthwhile but unremu-

[14] The distinction introduced in Simon (1978), although he uses it to describe the attitudes of different people with respect to a single value, namely 'utility'. See also **?**, Bench-Capon et al. (2012) and Atkinson and Bench-Capon (2016)

[15] The distinction is similar to that between *abstract unreachable goals* and *abstract goals* in Zurek (2017).

nerative, activity. This allows us to subsume Mercy under Generosity, while the qualification still acknowledges justice as a value.

The introduction of honesty may give a value ordering.

$L \succ H \succ C \succ G \succ D \succ J \succ P \succ F$

There is some scope for variation: e.g. P may be ranked higher than J without causing real problems to our moral vision. It is vital that honesty be given such a high ranking as there will normally be reasons based on some other value to break an agreement. Indeed it could be argued that H should even be preferred to L_s since it is always possible (and perhaps desirable) to avoid entering agreements which would risk demoting L_s

We might see a conflict between SN5a and SN2 and its relaxations SN2a and SN2b. In fact what we are doing is recognising a difference between those who cannot work, and whose requests should be granted, and those who could work but choose not to do so[16]. The distinction is intended to enforce SN1, but to allow for some excusable violations (e.g. on the grounds of illness). The notion that norms will, on occasion, need to be violated, and considered by many (e.g. Jones and Sergot (1992)) to be an essential feature of norms, will be the focus of later sections of this paper.

3.4.3 Turn Taking

In the previous subsection we considered situations with an initial imbalance of wealth. But it is possible, given recognition of a norm such as SN6, to enable the beneficial trade of surplus production for opportunities for $play_a$, through the mechanism of turn-taking. This arrangement, expressed in our model as one agent plays this year supported by another agent in return for supporting the play of that agent the following year, is in fact very common as an informal arrangement at the personal level. Many couples or groups living together will come to such an arrangement regarding chores, and the idea of "turn taking" is very common amongst children[17].

Turn taking also emerged in the empirical work of Lloyd-Kelly et al. (2014) in which a society of agents played a number of iterated prisoner's dilemma games. The agents had different degrees of *tolerance* (readiness to punish) and *responsiveness* (readiness to cooperate). What emerged was a number of stable situations: mutual cooperation and mutual defection, of course, but also some stable turn taking cycles. These turn taking cycles sometimes benefited the two agents equally, but even where one gained more from the arrangement than the other, it could still be beneficial to both, and to their combined score, when compared with mutual defection. Therefore we might well see such an arrangement emerge, even in an initially equal society, given that C is preferred to P and there is a reinforcing norm such as SN6. As has been noted above, such arrangements are likely to be especially common in domestic situations, where trust is likely to be high. This in turn suggests that it might be possible to differentiate H according to whom it is directed. It is not uncommon to regard it as wrong to cheat family and friends (H_f), dubious to cheat other individuals (H_i), but acceptable (where possible) to take advantage of large ("faceless") organisations

[16] The distinction between the deserving and undeserving poor was a central concern of the UK 1834 Poor Law Amendment Act, and is enjoying a revival in popular attitudes expressed in the UK today. It contrasts with the underlying philosophy of the UK Supplementary Benefits Act 1976, which saw a certain minimal level of support as the right of every citizen.

[17] Such informal turn taking is usually regulated by the peer group involved in the arrangement, and will not usually have any written agreements or legally enforceable status.

(H_o). Such discrimination is rarely enjoined by any ethical theory (although it is possible that, in some circumstances, it would be endorsed by some forms of consequentialism), but is a commonly argued for (and practiced) behaviour. Over-claiming on insurance is not uncommon and is seen by some as a "victimless" crime, suggesting that some might give H_o a very low rank, perhaps even below F.

3.4.4 Service Provision as Work

In several of the scenarios discussed previously it came about that because of the preference for C over some other values, certain agents may be enabled to $play_a$ because the consequent promotion of C was such that other agents were inclined to support this activity out of their surplus in preference to P. This is likely to be particularly so in the case of powerful agents who will choose to act as patrons to certain agents to allow and encourage certain kinds of $play_a$. But similar kinds of patronage may be attractive to other individuals as well, who may be prepared to part with a (typically) small part of their surplus to enable particular kinds of $play_a$ by particular agents. It is possible that this may emerge with just two agents. The ant may find the singing of the grasshopper so entertaining that he is willing to sacrifice his entire surplus for the privilege of listening to her sing. But, since the singing of a single grasshopper may entertain a whole colony of ants, it is even more attractive if the cost of supporting the grasshopper can be shared across a large number of individuals. Where this is so, a variety of entertainers can be supported, and other services performed. Money greatly assists this arrangement, and places it on a formal, contractual footing, so that it falls under SN6. As such we might expect the emergence of a service and entertainments sector, where some agents were able to adopt the role of providers of C promoting activities willingly supported by groups of other agents.

This is likely to be increasingly the case when productivity rises, so that workers generate larger surpluses. Now we can adjust our notions of the difference between $play_a$ and $play_d$. We can see $play_a$ as being non-work activities for which people are prepared to pay, and $play_d$ as non-work activities for which people are not prepared to pay. This will require consideration of the agent as well as the activity: people will pay to watch Lionel Messi play football, but no one will pay to watch me play football. We therefore combine norms SN1 and MN4a into a single norm:

SN1a It is obligatory to $play_a$ or to *work*.

This differs from MN4 because that norm was directed at only a subset of agents (the powerful), whereas SN1a can be seen as universal. Moreover while MN4 is generally not supported by laws, SN1a does lend itself to legislation. Interestingly, a norm like SN1a may be better supported by a system of reward for $play_a$ rather than punishment for $play_d$. Indeed the payment for the services provided for $play_a$ may well be seen in terms of reward for norm compliance. For a discussion of enforcing norms with rewards rather than punishments see Boer (2014).

3.4.5 Emergence of a State

As well as choosing to spend their surplus on providing themselves with culture, through paying others to $play_a$, agents may choose to pay others to do their duties. In Mahmoud et al. (2015) it was shown empirically that to avoid norms collapsing it is necessary that they not only be backed by the punishment of violators, but that those who fail to punish must themselves be punished. Since punishment can have a cost for the punisher, however, there are often reasons not to punish, and in societies where violations are comparatively rare, the cost of punishment can fall unevenly and unpredictably. We saw how punishment for violating SN1 can naturally be expressed as SN2 (which actually is cost free for the punisher in our model, but might involve a cost for the punisher in the general case), but when we move to more sophisticated norms such as SN6, recognising violations may not be straightforward. Recognising the need to punish is an important aspect of social cohesion: as expressed in Mahmoud et al. (2015):

> This move from enforcement by vigilantes (those taking the law into their own hands) to seeing law enforcement as the social duty of responsible citizens is an important milestone in the development of a society that respects its laws.

Once punishment is seen as a social duty it is a small step to organise and pay for a third party to punish violators. Assuming relatively few law breakers, a small levy will enable a dedicated agent to be paid to enforce the norms. Of course, non-payment of the levy will also be subject to punishment. From this it is a small step to taxation, and the provision of services such as law enforcement by the state, paid for from these taxes. And if law enforcement, why not other duties? Thus SN5a may be better observed by contribution to a central fund responsible for identifying those who should be supported and providing that support, than by individual charity.

In this way States may emerge, first as a Hobbesian *Leviathan* (Hobbes (1969)), but, once established, available to take on the performance of other duties. Further the State may take on the role of intervention to resolve conflicts of interest between its citizens (Chorley et al. (2006)), or to educate its citizens (Lloyd-Kelly et al. (2014)). An emergent State may also lead to new values such as *self-reliance*, *freedom* and *community*, and the relative preferences of these new values, and how they are promoted and demoted by different models of the State may provide insight into the form in which the State emerges. In some circumstances the State may take on an even broader role, and become itself the arbiter of what constitutes $play_a$, by itself supporting certain activities. Thus we often find subsidies for opera, but never for football. Of course, allowing the state to determine what counts as culture in this way will be controversial, and so we may find that we need to distinguish between two types of $play_a$: high culture as approved by the state and subsidised ($play_{sa}$) and popular culture approved by citizens and paid for out of their own retained surplus ($play_{pa}$). This provides another example of how increasing the level of sophistication of the model necessitates the finer grained discrimination of values and actions.

In the previous sections, we have discussed in general terms how one might equip agents to reason about norms and the values they serve, and the sort of norms that would correspond to such reasoning in various circumstances. We now turn to a specific use of these reasoning abilities in deciding when non-compliance with (i.e., violation of) norms can be justified, and, given there are often several ways in which norms can be violated, the form such violations should take.

4 Rules are Made to be Broken

4.1 Why Agents Might Need to Violate Norms

As observed in Hare (1965), for most people, most of the time, following norms involves little more than applying a set of learned principles. Hare, however, also says that there will be occasions when we need to think about a moral problem from first principles, and that the recognised norms are a useful way of encapsulating the conclusions of such reasoning for future use. This is a good way of viewing exceptions to existing norms: an existing norm needs to be violated, a way of violating it is chosen, and the key elements of the situation can then be captured and taken forward as an exception to the norm[18]. The working out of the automobile exception to the search provisions of the US Fourth Amendment provides an example of such a process in law (see Bench-Capon (2011)). Hare (1965) says:

> 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 crystalize 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.

We have already noted that an important reason for thinking in terms of norms is the recognition that on occasion they need to be violated (Jones and Sergot (1992)). While the norm is intended to provide a useful heuristic to guide behaviour, allowing for a quick unthinking response, unreflecting adherence to such guidelines is not what we we expect from a genuinely moral reasoner. Nor should it have been intended by the norm designer: the wise legislator will be well aware that not every eventuality can be foreseen[19], and that circumstances may arise in which the norm is, as Hamlet puts it, “more honoured in the breach than the observance”. While principles may serve well enough most of the time, there are situations in which the norm *should be violated* and in these situations we need to think through the situation from scratch.

The two main approaches to enforcing normative behaviour in MAS are either by removing prohibited actions (e.g. van der Hoek et al. (2007), Ågotnes et al. (2009), Dennis et al. (2016)) or by including explicit rules expressing the norms, akin to Asimov’s three principles of Robotics (Asimov (1950)). Often such principles are accompanied by sanctions. Neither are entirely satisfactory. The first gives no scope for violations, and while this may be adequate (even, by using model checking, such as Bošnački and Dams (1998), provably so) in situations covered by the model, in real life unexpected and unforeseen events will arise that take us outside the model. As for principles, at some point they are likely to give rise to conflicts (their main point as a literary device) or gaps, and unforeseen situations not covered by the principles will be encountered[20].

As an example of the unforeseen suppose that the ant has agreed to pay the grasshopper for entertaining him during the summer. When making the agreement the ant may have every intention of keeping the agreement, but 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 himself, 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

[18] The similarities with the classic four step *retrieve-reuse-revise-retain* model of case based reasoning in AI and presented in e.g. Aamodt and Plaza (1994) are clear.

[19] The advice to expect the unexpected has been given at least since Heraclitus (c.535 BC - 475 BC).

[20] That the “rules will run out” has been a key driver of AI and Law research since the days of Gardner (1984).

> 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.[21] Aquinas (2012), 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.

A common real world example in which actions are made unavailable is erecting bollards to prevent vehicles from entering a park (to use the famous example of Hart (2012)). What can be wrong with this approach? After all, we can *prove* that the undesirable situation the bollards are intended to prevent will not be reached, either using model checking (e.g. Bošnački and Dams (1998)) or analytic methods. Thus we can prove that universal compliance with the norm will achieve the desired results - as long as the situation is covered by the model. 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[22]. Similar considerations arise from obeying laws and principles: there is a conflict between the obligation to provide medical assistance and the obligation to exclude vehicles from the park. While any set of principles may provide good guidance most of the time, it is not difficult to think of situations 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 – removing or sanctioning prohibited actions – 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. 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. Value based reasoning offers the means to perform exactly the purpose driven reasoning required in these situations, which can explain both the general effectiveness of the norm (as illustrated in previous sections), and the need to violate it in some particular situations. We will, with an everyday traffic example, illustrate use of value based reasoning in deciding both to violate, and how to violate, norms. Before doing we so, we briefly review how rule breaking in AI and Law has been formalised in the literature.

4.2 Rule Breaking in AI and Law

That rules need to be broken should come as no surprise to anyone involved in AI and Law, or AI generally, although the issue is usually termed rule *defeasibility*, rather than

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

[22] This is known in economics as the *Theory of the Second Best* (e.g. Lipsey and Lancaster (1956)).

rule breaking or rule violation. In practice, very few rules do not permit exceptions, and not all exceptions are foreseeable, and so the exceptions cannot be represented *ex ante* as negative antecedents. Thus given the defeasible rule *if x is a bird x can fly*, there are many exceptions: flightless birds (e.g., penguins, kiwis and ostriches); a particular bird may be crippled, or temporarily unwell; the bird may have its feet set in concrete; the parrot may be dead, and so on. But, however many exceptions we list, it will always be possible to think of another one, and so some other way of handling the defeasibility is needed. Non-monotonic reasoning of this kind was perhaps the key concern of AI in the 80s and early 90s, and given the impracticality of listing all exceptions in rules' antecedents, a key concern was how to arbitrate between conflicting rules (e..g, the rule that most birds fly and the rule that most ostriches do not fly). A popular principle for adjudicating between conflicting rules was to prefer the more specific rule (e.g. Simari and Loui (1992)) so that the rule relating to ostriches is more specific, and so preferred to that relating to birds. In AI and Law, principles such as *Lex specialis derogat legi generali* (prefer the specific rule), *Lex posterior derogat legi priori* (prefer the later rule to earlier rules) and *Lex superior derogat legi inferiori* (prefer the rule originating from the superior source, national to local, Supreme Court to Appeal Court, etc) have been used (e.g., Valente (1995), and (Prakken (1991)). A general account of legal principles and their use in resolving normative conflicts can be found in Verheij et al. (1998). A final method of addressing conflicting rules was to specifically indicate which has priority either by relying on their order in the program code (e.g., Sergot et al. (1986)), or by the use of explicit priority rules (derived from precedent cases) as in Prakken and Sartor (1998).

Rarely, however, were AI and Law systems presented as coming to decisions: rather they were presented as proving the literal consequences of the law, as in Sergot et al. (1986), offering a suggestion which the user could accept or reject, or presenting both sides of the arguments for the user to choose between, as in HYPO (Ashley (1990)) and CATO (Aleven (1997)). Much of the work taken from general AI is directed towards theoretical reasoning, but legal decision making can be seen as practical reasoning, as argued for in Atkinson and Bench-Capon (2005). Giving a legal verdict can be seen (Al-Abdulkarim et al., 2016) as making a performative utterance, in the sense of Austin (1975), that is performing an action, rather than making a classification. Viewing legal decision making in this way sees the legal decision as a choice between two actions - decide for plaintiff or decide for defendant - and so brings it within the purview of value based practical reasoning as described above. The idea of value based reasoning in fact has its origins in law. The basic idea is that the strength of a rule depends on the values its promotes, and the preferences between values of its audience. That is, as explained in Bench-Capon (2003), that the acceptability of an argument depends on the values of its audience, an idea taken from the jurisprude Perelman (Perelman (1971) and Perelman (1980)). The notion was introduced into AI and Law in Berman and Hafner (1993), although there values were called *purposes*. Berman and Hafner argued that given conflicting precedents, one should follow the precedent which promoted the preferred purpose. This idea was developed in AI and Law, most fully in Bench-Capon and Sartor (2003) and evaluated empirically in Chorley and Bench-Capon (2005). Reasoning about the promotion and demotion of values was also formalised (in Deflog) in Verheij (2013).

4.3 Road Traffic Example

This section considers an aspect of everyday life where violation of the general norm is very common: the law that drivers should drive on the left (in the UK, on the right in many other

Table 1 Joint actions for self and on-coming in state 2100

self: on-coming	continue	slow	stop	change lane	mount pavement
continue	J1	J2	J3	J4	J5
slow	J6	J7	J8	J9	J10
stop	J11	J12	J13	J14	J15
change lane	J16	J17	J18	J19	J20
mount pavement	J21	J22	J23	J24	J25

countries). The law is intended to avoid collisions, and so promote the values of *Progress* and *Safety*. But on every journey, it is necessary to violate this law if progress is to be maintained: obstructions such as parked cars and roadworks, the need to overtake slow moving vehicles and cyclists and emergencies such as a pedestrian or animal stepping in front of the car, may all lead drivers to move to the right. But the question remains: when is it desirable to do so?

We will not give the full AATS here, since it would contain a lot of states and details which are not needed in our context, but we will give a sufficient fragment to allow the consideration of the relevant situations. Our AATS will represent relevant features of the states that can occur. For our example we consider:

- Whether there is an obstruction and whether it is moving or stationary (0 = clear, 1 = slow moving, 2 = stationary).
- Whether there is an on-coming vehicle and whether it can stop safely or not (0 = no on-coming, 1 = can stop safely, 2 = cannot stop).
- Whether our own vehicle can stop safely (0= can stop safely, 1 = cannot stop safely).
- Whether there has been[23] a collision (0 = no collision, 1(x,y) = x has collided with y).

For actions, both our own vehicle and the on-coming will be able to continue, change lane, stop, slow, or mount the pavement. For values we consider our own progress and safety (P(s) and S(s)), the progress and safety of the on-coming (P(o) and S(o)) and the safety of any pedestrians in the area (S(p)).

Now consider the transitions from the state where there is a stationary obstacle, and both ourselves and the on-coming could stop safely, and there has been no collision (2100). In this case there are a number of joint actions involving self and on-coming as shown in Table 1. Additionally, if one or both mount the pavement pedestrians may or may not be hit, depending on whether we get lucky or unlucky (often represented by including *Nature* in the joint action).

In the actions J1-J15 self obeys the norm, whereas J16-J20 and J21-J25 represent different ways of violating the norm. J1-J10 all result in self colliding with the obstacle, which demotes both P(s) and S(s). J11-15 do not demote S(s) but do demote P(s). Thus complying with the law will demote one or both of the values the norm was designed to promote, (although it does allow the on-coming to continue without threat to any of its values). We should therefore consider violating the norm. Suppose we go on to the right hand lane. Now J16 and J17 result in a head-on collision, which demotes all of P(s), P(o), S(s) and S(o). J18 demotes P(o) and J19 demotes both P(o) and S(o). J20 may or not demote S(p) for a variable number of pedestrians. Similarly J21-J25 will jeopardise the safety of an unknown number of pedestrians. We can therefore make a choice. If our desire for progress is insufficient to lead us to risk our safety (and the safety of others) we have to stop. If, however, we are sufficiently reckless that our desire for progress is such that we are willing to risk a collision we should change lane and hope that J18 is achieved, so that while P(o) is demoted, the threat

[23] The collision will occur during the transition.

to safety is avoided. This relies on the (normally acceptable) assumption that the on-coming agent will be less reckless than we are. J20-J25 are possible if we don't trust the on-coming to stop, but this poses the risk of an even more serious accident if we hit pedestrians. At this point we could either construct arguments for the other agents involved acting in certain ways (*the on-coming driver can't be as reckless as I am*, or *there will not be any pedestrians at this time of night*) in the manner of Atkinson and Bench-Capon (2007), or perform an expected value calculation as recommended in Atkinson and Bench-Capon (2016). Here most of us will choose to obey the norm. But if there is no on-coming, then we can change lane and violate the norm with no risk to safety. This will be better both than obeying the law or mounting the pavement, however unlikely we consider it to be that pedestrians are present. Thus value based reasoning can tell us both to violate the norm and how to violate it.

5 What Makes a Moral Audience?

As the discussion in section 4.3 showed, there may be more than one morally acceptable ordering on values: morality may dictate only a partial order. Some other orderings, such as one which licensed a refusal to pay the grasshopper even when her singing had been requested and there is a surplus available to do so, are not morally acceptable[24]. Also we have seen how different norms may evolve in different societies, driven by circumstance or by the preferences of its members. What we must do if we are to unleash autonomous agents on the world is to provide our agents with an acceptable ordering on which to base their reasoning, so that their choices will be acceptable to the societies in which they act. In order to do so, we need to look at the value order (total or partial) prevailing in that 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 Berman and Hafner (1993) and more formally presented in Bench-Capon and Sartor (2003). 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 (Chorley and Bench-Capon (2005)) the value ordering which best explains a set of cases was discovered by forming a theory to explain a set of cases, and then using the theory construction operators of Bench-Capon and Sartor (2003) to provide a better theory, in terms of explaining more cases, until the best available theory was found. In Bench-Capon et al. (2007), given a VAF 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. In that paper the ordering need not be fully determined (a *specific* audience): it is possible that the desired set of arguments can be accepted by several audiences, represented as a partial order on the values. In Modgil and Bench-Capon (2008), the VAF is rewritten as a meta-level argumentation framework (metalevel frameworks are more fully explored in Modgil and Bench-Capon (2010)), from which value orderings can emerge, or be 'formed',

[24] This is important if we are to accommodate the fact that different preferences are exhibited by different societies, or by the same society at different times, while avoiding the extreme relativist position of "anything goes".

as a result of dialogue games based on the rewritten frameworks. If desired explicit arguments for value orderings can be made using the Extended Argumentation Frameworks of Modgil (2009) which allow attacks to attack attacks as well as arguments.

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 in Bex et al. (2009). 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 Bex and Bench-Capon (2014) and Bex et al. (2014). 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 Wardeh et al. (2009)), 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, if the decision is made at the appropriate level by the appropriate authority, form important precedents, identifying apparent exceptions to existing norms, closing gaps and resolving conflicts, often revealing or choosing between value orderings as they do so.

As noted above, these methods may produce not a specific audience, but rather 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, in a multi-agent system, to impose the designer's view as to what is moral, and in a human society give rise to a homogeneous, conformist culture. Often, however, a degree of diversity may prove welcome, leading to different agents occupying different social roles. (cf. Durkheim (1893)'s notion that an advanced society promotes specialisation and division of labour. This is something that could be determined through future empirical investigation).

6 Discussion

When thinking about the foundations of norms, it is necessary to go beyond the norms themselves and think about their *rationales* in the particular societies in which they are found. We have argued that value-based practical reasoning applied to a model of society expressed as an AATS+V provides the machinery to model this kind of reasoning. Much current work on norm emergence and norm design is done using either simulations of public goods games or by proving properties of such games as in Shoham and Tennenholtz (1997), or by performing model checking on state transition diagrams as in Wooldridge and van der Hoek (2005). The first approach has given some insights, but the necessary simplifications, and assumptions about the homogeneity of agents, suggest that there are limitations to the approach. These doubts are strengthened by the fact that the behaviour of people observed empirically in experiments using such games does not support the model used (Engel (2011) for the Dictator game and Oosterbeek et al. (2004) for the Ultimatum game). The second approach also has a view of agents as highly goal directed, and tends to simplify its representation of norms by removing transitions representing forbidden actions. This means that it is highly effective at proving properties of the system, such as efficacy and liveness when the norms are complied with and for verifying the design of norms, but less good in explaining where the norms come from in the first place, and why the agents wish to pursue them. This is problematic if we are looking for a justification (other than a pragmatic one) so that the norm subjects can be persuaded rather than simply accept the norms imposed by a designer. We believe that the use of value-based argumentation provides a finer grained account of the reasoning involved, and is therefore better placed to account for the norms that emerge from different social set-ups and how they can be justified.

In section 3 we described how two norms might be justified by applying value based argumentation to a scenario modelling a simple society. One is a primary norm, while the other provides a natural way of punishing transgressions of the primary norm (and a way of removing transgressors). We believe that although the model is simple, it is a not implausible representation of a primitive agricultural society. Subsequently we described how making the model more sophisticated would lead to other norms, and more importantly to the need to introduce additional values (some of which may be *metavalues* promoted and demoted by value orderings rather than actions) and to make finer grained discriminations both in values and in actions. Thus *play* becomes seen as the socially beneficial $play_a$ and the indulgent $play_d$ and a need to discriminate the value of honesty according to the relationship between the agents involved in the transaction may become apparent. Unfortunately the provision of detailed models, and the particular arguments that they support, is beyond the scope of this paper (but see Atkinson and Bench-Capon (2016) for a discussion of how different value profiles lead to different behaviours in the Ultimatum game): all that was possible here is to sketch how additions to the model would result in different norms, and so give a flavour of the process.

We believe that such detailed models would indeed provide a fruitful way of analysing and explaining social developments. Our account here for example, coheres well with the account of social development found in Durkheim (1893). Durkheim suggests that in a "primitive" society people act and think alike with a collective or common conscience, which is what allows social order to be maintained. In such a society laws tend to be highly repressive. Both of these are true of the model presented in section 3, where there is a norm (SN1) to be followed by all and transgressions are effectively punished by death through SN2. Durkheim further argues that in an advanced, industrial, capitalist society, the complex division of labor means that people are allocated in society according to merit and rewarded accordingly, and that diversity is embraced rather than opposed. This accords with our discussion of the norms that develop as surplus production increases, and the development of exchanges enabled by SN6, leading to the increasing prevalence and diversity of service work, rather than food production. Within this framework we could, for example, explore the different norms that emerge when the surplus is due to a general rise in productivity, and when it is the result of an external boost to wealth, as in sixteenth century Spain. Note also that the sophisticated societies require increased cooperation (supported by norms such as SN6 and values such as Trust and Honesty) and tend to increase the degree of commercial exchanges between agents. It was these two factors that were found to lead to the greatest deviation from the classical model in the Ultimatum Games studied in Henrich et al. (2001), supporting the view that the more sophisticated the society the less adequate the model provided by simple public goods game simulations. Thus, even if simulations provide a good account of how initial norms *emerge*, investigating their *development* may require a finer grained approach.

7 Concluding Remarks

We have argued in this paper for the position that as agents display more and more autonomy and are starting (as, for example, drones and driverless cars) to interact more and more with the real world, they need to be given a capacity to reason explicitly with norms, in particular so that they can detect situations in which their norms need to be violated, and to choose the best way to violate them. While regimented systems and the rigid following of fixed rules might be adequate for agents working with limited autonomy in a closed environment, they are insufficient to allow agents to participate in the real world where they may be called upon to make life or death decisions. We have further argued that one way to provide this capacity is to make use of the value based practical reasoning techniques developed over the last decade. These techniques have the potential to explain why a particular society has the norms it does, and when and how these norms should be violated.

The literature also offers a number of approaches in which the value 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 Lloyd-Kelly et al. (2012) and Mahmoud et al. (2015). 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 Christie (2012)) and Phillipa Foot's famous *Trolley Problem* originally stated in Foot (2002) and recently revived in Bonnefon et al. (2016).

Future work will need to investigate several aspects of value based reasoning, including: inducing value orderings; consideration of the extents 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. We will also wish to use Extended Argumentation Frameworks to model the effect of making the value preferences explicit, and to enable different value preferences to be argued for.

Acknowledgements

The authors would like to acknowledge the comments of those who have reviewed this work at various stages of its development: for ECAI 2016, for the ECAI workshop AI4Justice, for Jurix 2016 and for this journal, and the editors of this special issue.

References

Aamodt, A. and Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1):39–59.

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

Ågotnes, T., van der Hoek, W., Tennenholtz, M., and Wooldridge, M. (2009). Power in normative systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 145–152. International Foundation for Autonomous Agents and Multiagent Systems.

Ågotnes, T. and Wooldridge, M. (2010). Optimal social laws. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1*, pages 667–674. International Foundation for Autonomous Agents and Multiagent Systems.

Al-Abdulkarim, L., Atkinson, K., and Bench-Capon, T. (2016). Statement types in legal argument. In *Proceedings of JURIX 2016*, pages 3–12.

Aleven, V. (1997). *Teaching case-based argumentation through a model and examples*. PhD thesis, University of Pittsburgh.

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

Aquinas, T. (2012). *Summa theologica (written 1265-74)*. Authentic Media Inc.

Aristotle (1962). *The Nicomachean ethics of Aristotle, translated by W.D. Ross*. Heinemann.

Ashley, K. (1990). *Modelling legal argument: Reasoning with cases and hypotheticals*. Bradford Books/MIT Press, Cambridge, MA.

Asimov, I. (1950). *I, Robot*. Robot series. Bantam Books.

Atkinson, K. and Bench-Capon, T. (2005). Legal case-based reasoning as practical reasoning. *Artificial Intelligence and Law*, 13(1):93–131.

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

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

Atkinson, K. and Bench-Capon, T. (2014). Taking the long view: Looking ahead in practical reasoning. In Parsons, S., Oren, N., Reed, C., and Cerutti, F., editors, *Computational Models of Argument - Proceedings of COMMA 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 109–120. IOS Press.

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

Atkinson, K. and Bench-Capon, T. (2016). Value based reasoning and the actions of others. In *22nd European Conference on Artificial Intelligence*, pages 680–688.

Atkinson, K., Bench-Capon, T., and Modgil, S. (2006). Argumentation for decision support. In Bressan, S., Küng, J., and Wagner, R., editors, *Proceedings of Seventeenth DEXA Conference*, volume 4080 of *Lecture Notes in Computer Science*, pages 822–831. Springer.

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review*, 80(04):1095–1111.

Axelrod, R. (1987). *The evolution of cooperation*. Basic Books, New York.

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

Bench-Capon, T. (2011). Relating values in a series of Supreme Court decisions. In Atkinson, K., editor, *Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth Annual Conference*, pages 13–22. IOS Press.

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

Bench-Capon, T. (2016a). Value-based reasoning and norms. In *22nd European Conference on Artificial Intelligence*, pages 1664–1665.

Bench-Capon, T. (2016b). Value-based reasoning and norms. In *Artificial Intelligence for Justice*, pages 9–17.

Bench-Capon, T., Atkinson, K., and Chorley, A. (2005). Persuasion and value in legal argument. *Journal of Logic and Computation*, 15(6):1075–1097.

Bench-Capon, T., Atkinson, K., and McBurney, P. (2012). Using argumentation to model agent decision making in economic experiments. *Autonomous Agents and Multi-Agent Systems*, 25(1):183–208.

Bench-Capon, T., Doutre, S., and Dunne, P. (2007). Audiences in argumentation frameworks. *Artificial Intelligence*, 171(1):42–71.

Bench-Capon, T. and Modgil, S. (2016a). Rules are made to be broken. In *Artificial Intelligence for Justice*, pages 18–21.

Bench-Capon, T. and Modgil, S. (2016b). When and how to violate norms. In Bex, F. and Villata:, S., editors, *Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Ninth Annual Conference*, pages 43–52. IOS Press, Amsterdam.

Bench-Capon, T., Prakken, H., and Visser, W. (2011). Argument schemes for two-phase democratic deliberation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, pages 21–30. ACM.

Bench-Capon, T. and Sartor, G. (2003). A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence*, 150(1):97–143.

Bentham, J. (1825). *The rationale of reward*. John and HL Hunt.

Berman, D. and Hafner, C. (1993). Representing teleological structure in case-based legal reasoning: the missing link. In *Proceedings of the 4th International Conference on Artificial Intelligence and Law*, pages 50–59. ACM.

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

Bex, F. and Bench-Capon, T. (2014). Understanding narratives with argumentation. In Parsons, S., Oren, N., Reed, C., and Cerutti, F., editors, *Computational Models of Argument - Proceedings of COMMA 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 11–18. IOS Press.

Bex, F., Bench-Capon, T., and Atkinson, K. (2009). Did he jump or was he pushed? *Artificial Intelligence and Law*, 17(2):79–99.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Binmore, K. (1998). Review of robert axelrod complexity and cooperation. *Journal of Artificial Societies and Social Simulation*, 1(1).

Boer, A. (2014). Punishments, rewards, and the production of evidence. In Hoekstra, R., editor, *Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Seventh Annual Conference*, pages 97–102. IOS Press, Amsterdam.

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576.

Bošnački, D. and Dams, D. (1998). Discrete-time promela and spin. In *Formal Techniques in Real-Time and Fault-Tolerant Systems*, pages 307–310. Springer.

Bratman, M. (1999). *Intention, Plans, and Practical Reason*. The David Hume Series. Cambridge University Press.

Broersen, J. (2014). Responsible intelligent systems. *KI-Künstliche Intelligenz*, 28(3):209–214.

Broersen, J., Dastani, M., Hulstijn, J., Huang, Z., and van der Torre, L. (2001). The BOID architecture: Conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the fifth international conference on Autonomous agents*, pages 9–16. ACM.

Burgemeestre, B., Hulstijn, J., and Tan, Y.-H. (2011). Value-based argumentation for justifying compliance. *Artificial Intelligence and Law*, 19(2-3):149–186.

Cartwright, D. and Atkinson, K. (2009). Using computational argumentation to support e-participation. *Intelligent Systems, IEEE*, 24(5):42–52.

Chorley, A. and Bench-Capon, T. (2005). An empirical investigation of reasoning with legal cases through theory construction and application. *Artificial Intelligence and Law*, 13(3-4):323–371.

Chorley, A., Bench-Capon, T., and McBurney, P. (2006). Automating argumentation for deliberation in cases of conflict of interest. In Dunne, P. and Bench-Capon, T., editors, *Computational Models of Argument - Proceedings of COMMA 2006*, volume 144 of *Frontiers in Artificial Intelligence and Applications*, pages 279–290. IOS Press.

Christie, G. (2012). *The notion of an ideal audience in legal argument*. Springer.

Dennet, D. (1995). *Darwin's Dangerous Idea*. Simon & Schuster.

Dennis, L. A., Fisher, M., Slavkovik, M., and Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14.

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

Dunne, P. (2010). Tractability in value-based argumentation. In Baroni, P., Cerutti, F., Giacomin, M., and Simari, G., editors, *Computational Models of Argument - Proceedings of COMMA 2010*, volume 216 of *Frontiers in Artificial Intelligence and Applications*, pages 195–206. IOS Press.

Durkheim, E. (2014. First published 1893). *The division of labor in society*. Simon and Schuster.

Engel, C. (2011). Dictator games: a meta study. *Experimental Economics*, 14(4):583–610.

Esteva, M., De La Cruz, D., and Sierra, C. (2002). Islander: an electronic institutions editor. In *Proceedings of First International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 1045–1052.

Foot, P. (2002). *Virtues and vices and other essays in moral philosophy*. CUP.

Garcez, A., Gabbay, D., and Lamb, L. (2005). Value-based argumentation frameworks as neural-symbolic learning systems. *Journal of Logic and Computation*, 15(6):1041–1058.

Gardner, A. v. d. L. (1984). *Artificial intelligence approach to legal reasoning*. MIT Press.

Governatori, G. (2015). Thou shalt is not you will. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 63–68. ACM.

Grasso, F., Cawsey, A., and Jones, R. (2000). Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition. *International Journal of Human-Computer Studies*, 53(6):1077–1115.

Hare, R. M. (1965). *Freedom and reason*. Oxford Paperbacks.

Hart, H. (2012). *The concept of law*. OUP Oxford.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001). In search of Homo Economicus: Behavioral experiments in 15 small-scale societies. *The American Economic Review*, 91(2):73–78.

Hobbes, T. (1969). *Leviathan, 1651*. Scolar Press.

J. F. Bonnefon, A. Shariff, I. R. (2016). The social dilemma of autonomous vehicles. *The Social Dilemma of Autonomous Vehicles*, 352(6293):41573–1576.

Jones, A. and Sergot, M. (1992). Deontic logic in the representation of law: Towards a methodology. *Artificial Intelligence and Law*, 1(1):45–64.

Kant, I. (1998. First Published 1785.). *Kant: Groundwork of the Metaphysics of Morals*. Cambridge Texts in the History of Philosophy. Cambridge University Press.

Levesque, H. J., Cohen, P. R., and Nunes, J. H. (1990). On acting together. In *Proceedings of the 8th National Conference on Artificial Intelligence*, pages 94–99.

Lipsey, R. and Lancaster, K. (1956). The general theory of second best. *The Review of Economic Studies*, 24(1):11–32.

Lloyd-Kelly, M., Atkinson, K., and Bench-Capon, T. (2012). Emotion as an enabler of co-operation. In *ICAART 2012 - Proceedings of the 4th International Conference on Agents and Artificial Intelligence*, pages 164–169.

Lloyd-Kelly, M., Atkinson, K., and Bench-Capon, T. (2014). Fostering co-operative behaviour through social intervention. In *Proceedings of (SIMULTECH), 2014*, pages 578–585. IEEE.

Loewenstein, G. (1999). Experimental economics from the vantage-point of behavioural economics. *The Economic Journal*, 109(453):25–34.

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

Mill, J. (1871). *Utilitarianism*. Longmans, Green, Reader, and Dyer.

Modgil, S. (2009). Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9):901–934.

Modgil, S. and Bench-Capon, T. (2008). Integrating object and meta-level value based argumentation. In Besnard, P., Doutre, S., and Hunter, A., editors, *Computational Models of Argument - Proceedings of COMMA 2008*, volume 172 of *Frontiers in Artificial Intelligence and Applications*, pages 240–251. IOS Press.

Modgil, S. and Bench-Capon, T. (2010). Metalevel argumentation. *Journal of Logic and Computation*, pages 959–1003.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21.

Nawwab, F., Bench-Capon, T., and P.Dunne (2008). A methodology for action-selection using value-based argumentation. In Besnard, P., Doutre, S., and Hunter, A., editors, *Computational Models of Argument - Proceedings of COMMA 2008*, volume 172 of *Frontiers in Artificial Intelligence and Applications*, pages 264–275. IOS Press.

Nofal, S., Atkinson, K., and Dunne, P. E. (2014). Algorithms for decision problems in argument systems under preferred semantics. *Artificial Intelligence*, 207:23–51.

Oosterbeek, H., Sloof, R., and Van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2):171–188.

Payne, T. and Tamma, V. (2015). Using preferences in negotiations over ontological correspondences. In *PRIMA 2015: Principles and Practice of Multi-Agent Systems*, pages 319–334. Springer.

Perelman, C. (1971). *The new rhetoric*. Springer.

Perelman, C. (1980). *Justice, law and argument: Essays on moral and legal reasoning*. D. Reidel/Kluwer Boston, MA.

Prakken, H. (1991). *A formal theory about preferring the most specific argument*. Vrije Universiteit, Faculteit der Wiskunde en Informatica.

Prakken, H. (2002). An exercise in formalising teleological case-based reasoning. *Artificial Intelligence and Law*, 10(1-3):113–133.

Prakken, H. and Sartor, G. (1998). Modelling reasoning with precedents in a formal dialogue game. *Artificial Intelligence and Law*, 6(2-4):231–287.

Rahwan, I. and Amgoud, L. (2006). An argumentation based approach for practical reasoning. In *Proceedings of AAMAS 06*, pages 347–354.

Rao, A. S. and Georgeff, M. (1991). Modeling rational agents within a BDI-architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Rep-*

resentation and Reasoning (KR'91), pages 473–484.

Rapoport, A. and Chammah, A. (1965). *Prisoner's dilemma: A study in conflict and cooperation*, volume 165. University of Michigan press.

Raz, J. (1979). *Practical Reasoning*. Oxford University Press, Oxford.

Roth, A. and Murnighan, J. K. (1978). Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical psychology*, 17(2):189–198.

Russell, S., Dewey, D., and Tegmark, M. (2016). Research priorities for robust and beneficial artificial intelligence: An open letter. *AI Magazine*, 36(4):3–4.

Sartor, G. (2010). Doing justice to rights and values: teleological reasoning and proportionality. *Artificial Intelligence and Law*, 18(2):175–215.

Savarimuthu, B., Purvis, M., Purvis, M., and Cranefield, S. (2008). Social norm emergence in virtual agent societies. In *Declarative Agent Languages and Technologies VI*, pages 18–28. Springer.

Sen, S. and Airiau, S. (2007). Emergence of norms through social learning. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1507–1512.

Sergot, M. J., Sadri, F., Kowalski, R. A., Kriwaczek, F., Hammond, P., and Cory, H. T. (1986). The British Nationality Act as a logic program. *Communications of the ACM*, 29(5):370–386.

Shoham, Y. and Tennenholtz, M. (1997). On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94(1):139–166.

Simari, G. R. and Loui, R. P. (1992). A mathematical treatment of defeasible reasoning and its implementation. *Artificial intelligence*, 53(2-3):125–157.

Simon, H. A. (1978). Rationality as process and as product of thought. *The American Economic Review*, 68(2):1–16.

Skyrms, B. (2014). *Evolution of the social contract*. Cambridge University Press.

Sugawara, T. (2011). Emergence and stability of social conventions in conflict situations. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 371–378.

Tremblay, J.and Abi-Zeid, I. (2016). Value-based argumentation for policy decision analysis: methodology and an exploratory case study of a hydroelectric project in Québec. *Annals of Operations Research*, 236(1):233–253.

Trojahn, C., Quaresma, P., and Vieira, R. (2008). An extended value-based argumentation framework for ontology mapping with confidence degrees. In Rahwan, I., Parsons, S., and Reed, C., editors, *Argumentation in Multi-Agent Systems 2007*, volume 4946 of *Lecture Notes in Computer Science*, pages 132–144. Springer.

Ullmann-Margalit, E. (1977). *The emergence of norms*. Clarendon Press Oxford.

Urmson, J. (1953). The interpretation of the moral philosophy of js mill. *The Philosophical Quarterly (1950-)*, 3(10):33–39.

Valente, A. (1995). *Legal Knowledge Engineering: A modelling approach*. IOS Press.

van Den Hoven, J. and Lokhorst, G, J. (2002). Deontic logic and computer-supported computer ethics. In Moor, J. and Bynum, T. W., editors, *Cyberphilosophy: The Intersection of Philosophy and Computing*, pages 376–386. Blackwell.

van der Hoek, W., Roberts, M., and Wooldridge, M. (2007). Social laws in alternating time: Effectiveness, feasibility, and synthesis. *Synthese*, 156(1):1–19.

van der Weide, T., Dignum, F., Meyer, J.-J. C., Prakken, H., and Vreeswijk, G. (2011). Multi-criteria argument selection in persuasion dialogues. In *Proceedings of the 10th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2011)*, pages 136–153.

Verheij, B. (2013). Arguments about values. In Atkinson, K., Prakken, H., and Wyner, A., editors, *From Knowledge Representation to Argumentation in AI, Law and Policy Making. A Festschrift in Honour of Trevor Bench-Capon on the Occasion of his 60th Birthday*, pages 243–257. College Publications.

Verheij, B. (2016). Formalizing value-guided argumentation for ethical systems design. *Artificial Intelligence and Law*, 24(4):387–407.

Verheij, B., Hage, J. C., and Van Den Herik, H. J. (1998). An integrated view on rules and principles. *Artificial Intelligence and Law*, 6(1):3–26.

Walker, A. and Wooldridge, M. (1995). Understanding the emergence of conventions in multi-agent systems. In *Proceedings of the First International Conference on Multi-agent Systems*, pages 384–389.

Wardeh, M., Bench-Capon, T., and Coenen, F. (2009). Padua: a protocol for argumentation dialogue using association rules. *Artificial Intelligence and Law*, 17(3):183–215.

Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley and Sons.

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

Zurek, T. (2017). Goals, values, and reasoning. *Expert Systems with Applications*, 71:442–456.