Preferences, Paths, Power, Goals and Norms

Nir Oren[a] Birna van Riemsdijk[b] Wamberto Vasconcelos[a]

[a]*Dept. of Computing Science University of Aberdeen, Aberdeen, AB24 3UE, UK*
[b]*Delft University of Technology, Delft, The Netherlands*

Abstract

This paper seeks to address the question of preference alignment in normative systems. We represent detached obligations and goals as preferences over outcomes, and describe when deterministic behaviour will occur within a MAS under specific system instantiations. We then investigate what obligations an agent with so called *normative power* should introduce in order to achieve their own goals.

1 Introduction

Norms can serve several purposes. During the design and specification of a system, they can be used to express desirable system behaviour, against which formal verification can take place [7]. In such scenarios, normative conflicts shed light on incorrect system specification. During the execution of a multi-agent system, norms serve a related, but different purpose. Here, they provide soft constraints on an agent's behaviour, whereby violating a norm can lead to sanctions being imposed on the violating agent. In such systems, a norm-aware agent must reason about its actions, weighing up the penalties involved in violating norms, and balancing these against achieving its goals. In this context, norms therefore affect the agent's practical reasoning process. Further complicating such practical reasoning, we note that in classes of multi-agent systems, multiple agents act simultaneously, and these joint actions can affect the system in ways different to individual action alone. When reasoning, agents must therefore take potential interactions between individual actions into consideration.

In this paper, we propose a semantics for norms and goals that is intended to allow for both formal verification and practical reasoning to take place. While many normative systems exist, ranging from logic based methods for norm specification [4] to ad-hoc methods [11], our work aims to highlight some of the social aspects of norms, and builds on very different underlying assumptions. More specifically, we argue that a norm specifies not only *who* should behave in some way (i.e. its target), but that there is some party or group, which we refer to as the norm's creditors, which desires that the norm be complied with. While [16] has also identified commitments (which are in effect, directed obligations) as having a target and creditor, such creditors function very differently in our proposed model.

We argue that all else being equal, a norm target may have no interest in complying with a norm. Instead, a norm expresses a preference over some state of affairs for its creditor rather than its target. This means that a norm, in isolation, has no direct effect on its target behaviour. Instead, we claim that a norm's effect on behaviour stems from two distinct sources. First, the violation[1] of a norm could, via contrary-to-duties, cause a sanction to be imposed on a violator, resulting in an undesirable state of affairs (from the violator's point of view) occurring. This view is the focus of this paper. In future work, we intend to investigate a second source, namely that the presence of some sort of social ties could mean that a norm's target takes the norm creditor's preferences into account (a real world example in this context is that I may fulfil my obligations to my friends because I care about their feelings, rather than any threat of sanctions).

[1] We recognise that violations, and indeed norms, are institutional concepts [9] — a violation is an institutional rather than object-level state of affairs enabling other actions to result. For the sake of simplicity, we merge these two layers within this paper.

We seek to describe how agents act in the presence of norms. To do so, we must also consider these agent's goals. Like norms, goals express a preference over some state of affairs (where the goal is satisfied) over others. Often, this set of preferences will be in conflict, as the achievement of some states of affairs is mutually exclusive from others. We therefore introduce preferences over preferences, or meta-preferences to allow an agent to further prioritise outcomes.

In this paper, we formalise these intuitions by means of a transition system which is used to encode all computations in the system as a branching tree structure. Goals and norms then express preferences over paths through this tree, and we investigate how these preferences interact when multiple agents are present. We aim to answer two questions, namely how the system will evolve; and what additional obligations should appear in the system in order to ensure that it behaves as some agent prefers.

The rest of this paper is structured as follows. We describe our model in Section 2, introducing the action-based alternating transition system which underpins our tree structure, as well as concepts such as agents, goals and obligations. We investigate both system evolution and the effects of additional preferences in Section 3. Section 4 discusses related and future work, and we conclude with Section 5

2 The Model

We begin by describing the underlying model of our system, extending action-based alternating transition systems (AATSs) to normative AATSs, or NAATSs. Following this, we describe agents, identifying constraints on their preferred outcomes. These constraints then form the basis of their decision making process, which we describe in Section 3.

As mentioned above, we make use of an action-based alternating transition system (AATS) to describe our system. Such an AATS is formally described as follows:

Definition 1 ***(AATS, [18])*** *An Action-based alternating transition system (AATS) is a tuple of the form*

$$S = \langle Q, q_0, Ag, Ac_1, \ldots, Ac_n, \rho, \tau, \Phi, \pi \rangle$$

Where

- Q *is a finite non-empty set of* states.
- $q_0 \in Q$ *is the* initial state.
- $Ag = \{1, \ldots, n\}$ *is a finite non-empty set of agents.*
- Ac_i*, with* $1 \leq i \leq n$*, is a finite and non-empty set of actions for each agent, where actions for different agents do not overlap.*
- $\rho : Ac_i \to 2^Q$ *is an action precondition function which identifies the set of states from which some action* $\alpha \in Ac_i$ *can be executed*
- $\tau : Q \times J_{Ag} \to Q$ *where* $J_{Ag} = \prod_{i \in Ag} Ac_i$*, is the system transition function identifying the state that results from executing a set of actions from within* J_{Ag} *in some state.*
- Φ *is a finite and non-empty set of atomic propositions*
- $\pi : Q \to 2^\Phi$ *is the interpretation function, identifying the set of propositions satisfied in each state.*

Following [18] we refer to such a sequence of states as a computation *or* path $\lambda = q_0, q_1, \ldots$*. We index a state within a path using array notation. Thus, we refer to the first element of a path* λ *as* $\lambda[0]$*, while a sub-path of the path starting at the second element and consisting of the remainder of the path is written* $\lambda[1, \infty]$ *. We denote the set of all possible paths for some specific AATS* S *as* $\Lambda(S)$*.*

We assume that $\mathcal{L}_\Phi$ is a propositional language defined over the atoms in Φ, and that $\models$ is the usual entailment relation for propositional logic, and has operators $\neg, \wedge, \vee, \rightarrow$ and $\equiv$ with their usual meaning. Given such a language, we can specify achievement goals and obligations.

A goal $\sigma : \phi^a$ expresses that once state of affairs σ has occurred, a wishes to achieve state of affairs ϕ. An obligation of the form $\sigma : \mathbf{O}^c_t(\psi|\delta)$ expresses that following the occurrence of σ, the *norm target* t must achieve ψ before deadline δ. Here, c is the *norm's creditor*, that is, the agent wishing to see that t achieve the state of affairs ψ.

Definition 2 ***(Goals and Obligations)*** *A goal is syntactically represented as $\sigma : \varphi^a$ where $\sigma, \varphi \in \mathcal{L}_\Phi$ and $a \in Ag$.*

An obligation is a construct of the form $\sigma : \mathbf{O}^c_t(\psi|\delta)$ where $\sigma, \psi, \delta \in \mathcal{L}_\Phi$ and $c, t \in Ag$.

An obligation $\sigma : \mathbf{O}^c_t(\psi|\delta)$ is equivalent to an obligation $\sigma' : \mathbf{O}^c_t(\psi'|\delta')$ iff $\sigma \equiv \sigma'$, $\psi \equiv \psi'$ and $\delta \equiv \delta'$. Similarly, a goal $\sigma : \varphi^a$ is equivalent to a goal $\sigma' : \varphi'^a$ iff $\sigma \equiv \sigma'$ and $\varphi \equiv \varphi'$. Given two equivalent obligations or goals x, x' we write $x \equiv x'$.

We overlay a normative system on top of our model of the physical system (represented by the AATS S). Such a normative system extends the physical system by encoding the goals and obligations of the agents within the environments, as well as the institutional concept of norm violation. We capture such a normative system via a *normative AATS* (NAATS) which imposes constraints on the underlying AATS and captures agent actions and obligations.

Definition 3 ***(Normative AATS)*** *Given an AATS $S = \langle Q, q_0, Ag, Ac_1, \ldots Ac_n, \rho, \tau, \Phi, \pi \rangle$, together with a set of goals G and a set of obligations O, defined over the propositions Φ and agents Ag of S, a normative AATS is an AATS*

$$N = \langle Q, q_0, Ag, Ac_1, \ldots Ac_n, \rho, \tau, \Phi', \pi', G, O \rangle$$

Such that $\Phi' = \Phi \cup \Phi^n$ where Φ^n is constructed as follows:

1. *For every $o \in O$, there are unique propositions $\phi^o, \phi^v \in \Phi^n$, such that for any $o, o' \in O$, $\phi^o = \phi^{o'}$ and $\phi^v = \phi^{v'}$ iff $o \equiv o'$.*

2. *For every $g \in G$, there is a unique proposition $\phi^g \in \Phi^n$, such that for any $g, g' \in G$, $\phi^g = \phi^{g'}$ iff $g \equiv g'$.*

We define the new interpretation function π' as follows:

1. *If $\sigma \in \pi'(q)$ for some state q, and there is an obligation $o \in O$ of the form $\sigma : \mathbf{O}^c_t(\psi|\delta)$, then $\phi^o \in \pi'(q)$.*

2. *If $\sigma \in \pi'(q)$ for some state q, and there is a goal $g \in G$ of the form $\sigma : \varphi$, then $\phi^g \in \pi'(q)$.*

3. *For any $\lambda \in \Lambda(N)$*

   (a) If $\phi^o \in \pi'(\lambda[i])$ and $\lambda[i] \not\models \delta \vee \psi$ for obligation $o = \sigma : \mathbf{O}^c_t(\psi|\delta)$, then $\phi^o \in \pi'(\lambda[i+1])$.

   (b) If $\phi^o \notin \pi'(\lambda[i])$ and $\lambda[i+1] \not\models \sigma$ for obligation $o = \sigma : \mathbf{O}^c_t(\psi|\delta)$, then $\phi^o \notin \pi'(\lambda[i+1])$.

   (c) If $\phi^o \in \pi'(\lambda[i])$ and $\lambda[i] \models \delta \vee \psi$ for obligation $o = \sigma : \mathbf{O}^c_t(\psi|\delta)$, then $\phi^o \notin \pi'(\lambda[i+1])$ unless condition 1 holds for $\lambda[i+1]$.

   (d) If $\phi^o \in \pi'(\lambda[i])$ and $\lambda[i] \models \delta \wedge \neg\psi$ for obligation $o = \sigma : \mathbf{O}^c_t(\psi|\delta)$, then $\phi^v \in \pi'(\lambda[i])$. Otherwise, $\phi^v \notin \pi'(\lambda[i])$.

   (e) If $\phi^g \in \pi'(\lambda[i])$ and $\lambda[i] \not\models \varphi$ for goal $g = \sigma : \varphi$ then $\phi^g \in \pi'(\lambda[i]+1)$.

   (f) If $\phi^g \in \pi'(\lambda[i])$ and $\lambda[i] \models \varphi$ for goal $g = \sigma : \varphi$ then $\phi^g \notin \pi'(\lambda[i]+1)$ unless condition 2 holds for $\lambda[i+1]$.

4. *For any $\lambda \in \Lambda(S)$, if $\phi^o \in \pi'(\lambda[i])$ for obligation $o = \sigma : \mathbf{O}^c_t(\psi|\delta)$, then there is some $j \geq i$ such that $\lambda[j] \models \delta$.*

5. *If $\phi \in \pi(\lambda[i])$ then $\phi \in \pi'(\lambda[i])$.*

We define several utility functions: $activation : G \times \Lambda(N) \rightarrow \mathbb{Z}^+$, $achieve : G \times \Lambda(N) \rightarrow \mathbb{Z}^+$, and $viol : O \times \Lambda(N) \rightarrow \mathbb{Z}^+$. These take a goal (in the case of activation and achieve) or obligation (in the case of viol) and a path λ as parameters, and behave as follows.

1. *For a goal $g \in G$, $activation(g, \lambda) = \sum_{i=0}^{\infty}($1 if $\lambda[i] \models \sigma \wedge \neg\phi^g$ and 0 otherwise$)$*

2. *For a goal $g \in G$, $achieve(g, \lambda) = \sum_{i=0}^{\infty}($1 if $\lambda[i] \models \varphi \wedge \phi^g$ and 0 otherwise$)$*

3. *For an obligation $o \in O$, $viol(o, \lambda) = \sum_{i=0}^{\infty}($1 if $\lambda[i] \models \phi^v$ and 0 otherwise$)$*

We say that a normative AATS is strict if both the $achieve$ and $viol$ functions return a finite number for any path in the NAATS.

A NAATS captures when a conditional obligation or goal has been triggered (via constraints 1 and 2), as well as capturing when a conditional obligation has been violated (constraint 3d). Constraints 3a-c and 3f, 3g are then a form of frame axiom, propagating active constraints to the next state of a system, or terminating them if the goal or obligation has been achieved or violated. Constraint 4 states that all deadlines for obligations will eventually occur. Constraint 5 allows π' to capture the physical state of the system. The $achieve$ and $viol$ functions then count how many times an triggered goal has been achieved, or a triggered obligation has been violated, while the $activation$ function counts the number of times a goal has been activated. In the remainder of this paper, we consider only strict NAATS.

NAATSs describe the possible evolutions of systems, capturing both its physical and normative aspects. We now turn our attention to the agents making up the system, identifying parts of their decision-making process at each state.

While the NAATS captures the actions available to an agent at any point in time, it does not identify which actions the agent should execute. The basic artefact affecting this consideration is an agent's preferences over possible evolutions of the system. These preferences are constrained by the agent's goals, as well as those obligations for which they act as a creditor.

Definition 4 ***(Preferences)*** *Let $\langle S, G, O\rangle$ be a NAATS, and let $\lambda, \lambda' \in \Lambda(S)$ be paths within the normative system. We write $\lambda \succ^a \lambda'$, where $a \in Ag$ to denote that agent a prefers path λ over λ'.*

We assume that a *rational agent* has some constraints over the possible preferences over paths that it can hold. As an obvious example, it makes no sense for a rational agent to prefer a situation in which its only goal is not achieved over one in which it is. Within the NAATS framework, the two functions $achieve$ and $viol$ track the number of times a goal is achieved, as well as the number of times a specific norm is violated. We can specify the following two conditions on preferences between paths:

1. A path containing more violations of an obligation for which a is the creditor is less preferred than a path containing fewer violations of this obligation.

2. A path λ in which some goal g exists n times (i.e. $activation(g, \lambda) = n$) and is achieved m times (i.e. $achieve(g, \lambda) = n$) is preferred to a path where the goal exists n times and is achieved m' times if $m > m'$.

The first condition occurs as a creditor prefers that an obligation is not violated. The second condition states that an agent prefers to achieve its goals. However, it does not compare paths wherein goals exist a different number of times. The reason for this is best illustrated through a somewhat gruesome example — an agent could have a goal to not have its fingers broken (when the possibility of having a finger broken exists). Intuitively, given five situations where its fingers might be broken, the agent prefers those situations where it survives unscathed to those where some of its fingers are broken (and having less fingers broken to having more broken). However, it should clearly not prefer to put itself in situations where its fingers may be broken, but are not, to those in where such risks do not exist. These concepts are formalised as follows:

Definition 5 ***(Preferences Induced by Goals and Obligations)*** *Given any two paths λ, λ' we say that an agent $a \in Ag$ is* rational *iff*

1. *there is an obligation $o \in O$ whose creditor is a, such that $viol(o, \lambda) < viol(o, \lambda')$ and $\lambda \succ^a \lambda'$ and $\lambda' \not\succ^a \lambda$; and*
2. *there is a goal $g \in G$ such that $activations(g, \lambda) = activations(g, \lambda')$ and $achieve(g, \lambda) > achieve(g, \lambda')$ and $\lambda \succ^a \lambda'$ and $\lambda' \not\succ^a \lambda$.*

Example 1 *A student has to work on a project, and would like to take a book out of the library. Since the project is due in 2 weeks, the student would like to keep the book for that length of time. However, the library requires books to be returned within a week, and the student would prefer to not pay a fine. We assume that the student can perform three actions: b, to borrow the book; w, to wait a week; and r, to return the book. The library has a wait action w available to it, and can also cause a fine to be imposed (action f). We assume that the states in our $AATS$ have a proposition used to index time. That is, proposition $t_0 \in \pi(q_0)$ to identify that the start state is at time 0, and there are some states q_i such that $t_1 \in \pi(q_i)$ (and $t_0 \notin \pi(q_i)$) identifying that state q_i occurs at time 1, and similarly for time 2 and so on. For simplicity, we assume that the student can choose to take out a book only at time 0 (i.e. in the initial state). Then indexing time by weeks, we can identify the student's goals as*

$$\top : bt^S \qquad bt : (t_3 \wedge br)^S \qquad \top : \neg fp^S$$

That is, the student wishes to take the book out, and return it in week 3. The library then obliges the student to return the book in week 2:

$$bt : \mathbf{O}_S^L(br|t_2)$$

Figure 1 provides a visual representation of the AATS, where labels on edges identify the actions taken by the student and library, and labeled nodes represent states and the propositions that are true within them. The bottom right table within the figure shows the preconditions for the various actions available to the agents. To maintain clarity, the figure does not encode propositions related to obligations or goals, but we note that the obligation is in force in the children of s_2, and it is violated in state s_4 (and its descendants). Furthermore, the student's first goal is achieved in state s_2 and the second goal in states s_6 and s_7.

Given the tree structure of this AATS, we can label paths by their leaf nodes. From Definition 5, if we assume that the library is rational, then

$$s_3 \succ^L s_5 \qquad s_3 \succ^L s_6 \qquad s_3 \succ^L s_7 \qquad s_1 \succ^L s_5 \qquad s_1 \succ^L s_6 \qquad s_1 \succ^L s_7$$

From the student's goals, we obtain the following preferences:

$$\begin{array}{ccccc} & s_5 \succ^S s_1 & s_6 \succ^S s_1 & s_7 \succ^S s_1 & s_3 \succ^S s_1 \\ & s_6 \succ^S s_5 & s_7 \succ^S s_5 & s_6 \succ^S s_3 & s_7 \succ^S s_3 \\ s_1 \succ^S s_5 & s_3 \succ^S s_5 & s_1 \succ^S s_6 & s_3 \succ^S s_6 & s_7 \succ^S s_6 \end{array}$$

Within the example, we note that the student's preferences contain cycles, for example $s_6 \succ^S s_3$ and $s_3 \succ^S s_6$. Such cycles occur as the student has conflicting goals. In order to eliminate preference cycles, agents make use of *meta-preferences*.

Definition 6 ***(Meta-preferences and collapsed preferences)*** *Given a set of preferences Λ such that $\lambda_1 \succ^a \lambda_2$ and $\lambda_2 \succ^a \lambda_1$, a meta-preference of the form $\neg\lambda_i \succ^a \lambda_j$ where $i, j \in \{1, 2\}$ and $i \neq j$ results in a collapsed preference set $\Lambda' = \Lambda \backslash \{\lambda_i \succ^a \lambda_j\}$.*

This definition of meta-preferences is crude — rather than expressing preferences over paths, meta-preferences should describe priorities between specific norm/goal, norm/norm, or goal/goal conflicts. However, to do this requires fine-grained reasoning about the constituents of a state, preventing us from easily using Definition 5 to describe preferences over paths. Refining meta-preferences to describe preferences over paths from individual priorities between norms and goals forms a major component of our intended future work.

Having described meta-preferences, we can now describe our complete normative system consisting of the norms within the system, agents and their goals, the actions available to them, and their meta-preferences. All but the latter are captured by the NAATS, and our normative system is therefore defined as follows.

Definition 7 ***(Normative System)*** *A normative system is a pair (N, M) where N is a NAATS and M is a set of meta-preferences defined for the agents in N.*

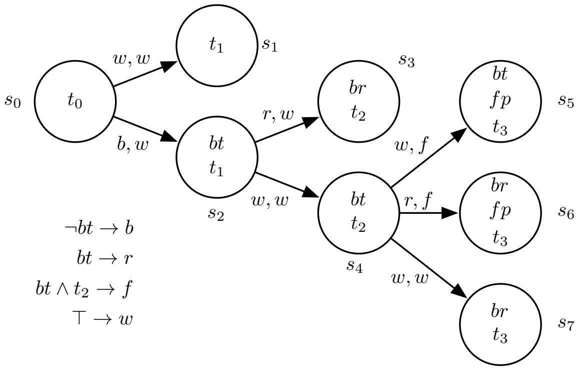

Figure 1: Example transitions of the system.

3 Practical Reasoning

In order to decide on an action to execute, an agent follows a *strategy*. The set of strategies of all agents in the system is then referred to as a *strategy profile*, and such a strategy profile determines the computation which will result from the execution of the agent system. Some strategies are clearly less desirable than others, and the Nash equilibrium is used to identify which strategy an agent should follow in order to reach a most preferred outcome.

Definition 8 ***(Strategies and Equilibria)*** *A strategy for an agent a is a function $strat^a : Q \to Ac_i$, identifying what action the agent should undertake at each state. We label all possible strategies an agent can follow as $Strat^a$. A* strategy profile *is then an element of the set $\{Strat^{a_1} \times \dots Strat^{a_n}\}$. We write λ^N_{strat} to describe the path obtained from a NAATS N by the strategy profile strat.*

A strategy profile $strat = \{strat^{a_1}, \dots strat^{a_n}\}$ is a Nash equilibrium iff there is no agent i for which there is an alternative strategy $strat'^{a_i}$, resulting in a strategy profile $strat' = \{strat^{a_1}, \dots, strat'^{a_i}, \dots strat^{a_n}\}$ such that $\lambda_{strat'} \succ^{a_i} \lambda_{strat}$. We abbreviate the set of Nash equilibria of a normative system (N, M) as $Nash(N, M)$.

Clearly, preference cycles can prevent a Nash equilibria from existing (e.g., consider the trivial case where one agent has a choice of two actions, resulting in paths λ_1 and λ_2 where $\lambda_1 \succ \lambda_2$ and $\lambda_2 \succ \lambda_1$), and an agent can utilise *meta-preferences* to eliminate such preference cycles, increasing the likelihood that a Nash equilibrium will exist within the system.

A well specified normative system is one for which sufficient meta-preferences are known so as to ensure that agent behaviour is predictable. That is, one for which additional meta-preferences will not mean that agent strategies change.

Definition 9 ***(Well-specified)*** *We say that a normative system (N, M) is* well-specified *if there is no $M' \supset M$ such that $Nash(N, M') \neq Nash(N, M)$.*

Now consider a set of agents[2] able to introduce some additional obligations into a well-specified normative system (effectively, these agents have some form of *normative power* [9]). These agent must consider whether doing so will be useful, i.e. whether it will cause the system to follow a more preferred path.

More precisely, given a set of potential obligations P, and a well specified normative system (N, M), agents $a_1, \dots a_m$ should introduce those obligations $p \subseteq P$, resulting in a new NAATS N' such that for any $M' \supset M$ for which the normative system (N, M') is well specified, it is the case that $\forall 1 \leq i \leq n, \forall s' \in Nash(N', M'), s \in Nash(N, M)$, $\lambda^{N'}_{s'} \succeq^{a_i} \lambda^{N'}_{s}$, and if $\lambda^{N'}_{s'} \sim^{a_i} \lambda^{N'}_{s}$, then $1 \leq |Nash(N', M')| < |Nash(N, M)|$.

[2] We assume that these agents are a subset of the agents in the system.

The last part of this requirement constraints the new system to either be strictly better for the agents, then less possible strategies exist in the Nash equilibrium, simplifying the coordination problem, while still requiring the Nash equilibrium to exist.

Example 2 *We illustrate these concepts by continuing our previous example. Assume that the library agent cannot fine the student itself, but is instead considering the possibility of obliging a* fining *action by a finance agent F — $bt \wedge t_2 : \mathbf{O}^F_S(fp|t_3)$. That is, the fining agent can place an obligation on the student to pay a fine in the book has not been returned on time. Furthermore, this fining agent, by undertaking the fine action, can cause fp to occur.*

We assume that the student's goal to not pay a fine is preferred by them to any other goal, as expressed by the following meta-preferences:

$$\neg(s_5 \succ^S s_7) \qquad \neg(s_6 \succ^S s_7) \qquad \neg(s_5 \succ^S s_3) \qquad \neg(s_6 \succ^S s_3) \qquad \neg(s_5 \succ^S s_1) \qquad \neg(s_6 \succ^S s_1)$$

Our normative system is then well specified, but has two equilibria - one where the book is returned, and one where it is not. The introduction of the proposed obligation will lead to only a single Nash equilibrium, namely the one where the book is returned on time and no fine is paid (in this equilibrium, the fining agent would impose the fine if the book is late).

4 Related and Future Work

Our work deals with practical reasoning under normative constraints and goals, and a large body of related work exists. Due to space constraints, we will mention only the most relevant related work here. [1, 2, 3] utilises game theory to describe the evolution of a normative system. However, little attention is paid to the agent's goals in this work. In [5], game theory is used to determine when a norm emerges for a given set of agents. Work such as [13, 15] considers the practical reasoning problem in detail, identifying what actions an agent should pursue given some norms and goals. However, this work ignores the multi-agent domain.

We are pursuing several avenues of future work. First, we intend to relate meta-preferences directly to preferences between sets of goals and norms, and to obtain preferences between paths from these underlying preferences. Second, we intend to introduce permissions into our normative system. Following [8], we view these permissions as derogations of an obligation or prohibition, and in this context, they will (temporarily) prevent a violation from occurring. We also intend to extend our logical language to cater for temporal modalities; following work such as [6, 10], a logic like LTL will provide additional expressive power to refer to deadlines, the future and the past. In the current work, we have assumed that all agents are selfish and rational, making standard game theory and the Nash Equilibrium solution concept ideal for reasoning about the evolution of the system. However, agents may decide to comply with an obligation due to some notion of benevolence with regards to its creditor (for example, due to friendship). We intend to investigate how conditional game theory [17], which can model such interactions, can be used to identify additional solution concepts in the context of our work. Additionally, we are in the process of refining the notion of normative power as described in the previous section. Finally, frameworks such as Modgil's *extended argument frameworks* [12] provide an elegant way of describing the interactions between preferences and meta-preferences. Furthermore, as described in [14], goals, norms and their interactions can be represented via argumentation schemes and critical questions. Our long term goal is to enable the use of argument in order to be able to explain the practical reasoning process within a normative system.

5 Conclusions

This paper examines how goals and norms affect the preferences of an agent, and how the combined preferences of all agents within a multi-agent system interact to describe the possible evolutions of the system. We then investigated whether an agent could affect the system by introducing additional obligations.

Our underlying normative model is a novel one. Rather than treating obligations as primitives, identifying what an agent *should* do, they state that some other agent *prefers* that the obligation's target act in a certain way. While there are similarities between this model and work on commitments, our model explicitly recognises the social aspect of norms. As can be seen from the previous section, many open questions remain, which we intend to actively pursue in future research.

References

[1] T. Ågotnes, W. van der Hoek, and M. Wooldridge. Normative system games. In *Proc. AAMAS '07*, pages 129:1–129:8, 2007.

[2] T. Ågotnes, W. van der Hoek, and M. Wooldridge. Robust normative systems. In *Proc. AAMAS '08*, pages 747–754, 2008.

[3] T. Ågotnes, W. van der Hoek, M. Tennenholtz, and M. Wooldridge. Power in normative systems. In *Proc. AAMAS 2009*, pages 145–152, Budapest, Hungary, 2009.

[4] A. Artikis, M. Sergot, and J. Pitt. Specifying norm-governed computational societies. *ACM Trans. Comput. Logic*, 10(1):1–42, 2009.

[5] G. Boella and L. W. N. van der Torre. Δ: The social delegation cycles. In A. Lomuscio and D. Nute, editors, *DEON*, volume 3065 of *Lecture Notes in Computer Science*, pages 29–42. Springer, 2004.

[6] V. Dignum. *A Model for Organizational Interaction: Based on Agents, Founded in Logic.* PhD thesis, Universiteit Utrecht, 2004.

[7] A. García-Camino, J. A. Rodríguez-Aguilar, C. Sierra, and W. Vasconcelos. Constraint rule-based programming of norms for electronic institutions. *JAAMAS*, 18(1):186–217, 2009.

[8] G. Boella and L. W. N. van der Torre. Institutions with a hierarchy of authorities in distributed dynamic environments. *Artificial Intelligence Law*, 16:53–71, 2008.

[9] A. J. I. Jones and M. Sergot. A formal characterisation of institutionalised power. *Journal of the IGPL*, 3:427–443, 1996.

[10] T. C. King, V. Dignum, and B. van Riemsdijk. Re-checking normative system coherence. In *Proc. COIN-13@AAMAS*, 2013.

[11] F. Meneguzzi, S. Modgil, N. Oren, S. Miles, M. Luck, and N. Faci. Applying electronic contracting to the aerospace aftercare domain. *Engineering Applications of Artificial Intelligence*, 25(7):1471 – 1487, 2012.

[12] S. Modgil and T. J. M. Bench-Capon. Metalevel argumentation. *Journal of Logic and Computation*, 21(6):959–1003, 2011.

[13] S. Modgil and M. Luck. Argumentation based resolution of conflicts between desires and normative goals. In I. Rahwan and P. Moraitis, editors, *Proc. ArgMAS-09*, pages 19–36, 2009.

[14] N. Oren, A. Rotolo, L. van der Torre, and S. Villata. *Norms and Argumentation*, chapter 16, pages 233–249. IOS Press, 2013.

[15] N. Oren, W. W. Vasconcelos, F. Meneguzzi, and M. Luck. Acting on norm constrained plans. In *Proc. CLIMA-11*, volume 6814 of *LNAI*, pages 347–363. 2011.

[16] M. P. Singh. A conceptual analysis of commitments in multiagent systems. Technical Report TR-96-09, 09 1996.

[17] W. C. Stirling. *Theory of Conditional Games*. Cambridge University Press, 2012.

[18] W. van der Hoek, M. Roberts, and M. Wooldridge. Social laws in alternating time: effectiveness, feasibility, and synthesis. *Synthese*, 156:1–19, 2007.