Towards socially sophisticated BDI agents

F. Dignum
Department of Mathematics and Computer Science
Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands
dignum@win.tue.nl

D. Morley
Department of Computer Science and Software Engineering
The University of Melbourne, Parkville, Victoria 3052, Australia
{dnm}@cs.mu.oz.au

E. A. Sonenberg
Department of Information Systems
The University of Melbourne, Parkville, Victoria 3052, Australia
{l.sonenberg}@dis.unimelb.edu.au

L. Cavedon
Department of Computer Science
RMIT University, Melbourne, Victoria 3001, Australia
cavedon@cs.rmit.edu.au

Abstract

We present an approach to social reasoning that integrates prior work on norms and obligations with the BDI approach to agent architectures. Norms and obligations can be used to increase the eficiency of agent reasoning, and their explicit representation supports reasoning about a wide range of behaviour types in a single framework. We propose a modified BDI interpreter loop that takes norms and obligations into account in an agent's deliberation.

1 Introduction

Many researchers have argued that in the design of multi-agent systems, to support rich collaborative behaviour it is essential to provide individual agents with various forms of social awareness. Technical concepts related to notions of social awareness that have been considered include: joint and shared plans [8, 13, 18], conventions and social responsibility [10, 11], social commitment [2] social laws [16], spheres of commitment [17], reasoning with obligations [1], etcetera. In this paper we present a new approach to social reasoning which integrates prior work on norms and obligations [3, 5, 20] with the now standard BDI approach to agent architectures [14].

In the BDI approach, the behaviour of an individual agent is shaped by the agent's state of knowledge about the environment (*beliefs*), the states of the world it seeks to bring about (*goals*), and the execution of pre-programmed (partial) *plans* that have been designed to bring about certain world states, in pre-specified circumstances. The plans selected by the agent for execution at any one time are referred to as *intentions*. Employing the BDI approach in *multi-agent systems* involves extending the agent's representation capability to support reasoning explicitly about the beliefs, goals, and perhaps intentions of other agents in the system.

The BDI approach has proved valuable for the design of agents that operate in a dynamic environment, and that operate flexibly and appropriately to changing circumstance despite incomplete information about the state of the world and other agents in it. The basic BDI deliberation cycle provides for a spectrum of behaviours ranging from purely deliberative, to highly reactive, depending on the structure of the plans provided, and on decisions such as how often to take account of changes in the environment while executing plans. A strength of the BDI approach is the opportunity for agent designers to build agents of different behaviours along

this spectrum within a single agent architecture. Here, we take steps towards enhancing the flexibility of the reasoning by incorporating social influences, expressed in terms of norms and obligations, allowing a rich spectrum of social behaviours to be described in a single framework.

We propose that both norms and obligations should be explicitly used as influences on an agent's behaviour. Norms, as present in human societies, assist in standardising the behaviour of individuals, making it easier to cooperate and/or interact within that society. The same holds for agent societies. Because agents in the society are designed so they tend to follow norms, knowledge of the norms allows for easier coordination, as certain behaviours of others can be anticipated, with some degree of reliability. Obligations[1], on the other hand, are associated with specific enforcement strategies which involve punishment of violators. In that sense, obligations are explicit tools to influence the behaviour of other (autonomous) agents and provide some stability and reliability in the interactions of agents, while allowing some flexibility. They provide a level of "freedom of choice" with explicit consequences on those choices. The main reason to make this distinction is that norms and obligations influence the behaviour of agents in a different way, as discussed further below.

A key part of our argument is that it is essential to allow explicit reasoning about norms and obligations. There are two reasons not to have norms and obligations hardwired into the agents. First, circumstances might change, which makes norms obsolete or suggests modified norms. A second reason not to hardwire the norm into the agents is if they interact with agents from other systems that follow different norms, explicit representation of norms and obligations can support appropriate, more flexible, reasoning.

The deliberative process of a standard BDI agent involves successively monitoring the incoming event stream, identifying changes in the environment to be responded to, generating goals, selecting (one or more) goals to pursue as intentions, and commencing/continuing execution of adopted plans, while monitoring. This abstract account of the deliberation cycle is of course silent on how the agent should identify urgent or important goals ahead of others, and implemented systems employ various strategies of numerical weightings, utilities, preference orderings, or other forms of meta-level reasoning to support the selection process. Here we propose the introduction of norms and obligations to support the socially motivated deliberation process of the agent. The socially sophisticated agents we deal with have explicit knowledge about the enacted norms in a multi-agent environment and make choices whether or not to obey norms, and how to weigh up the impact of punishments for obligation violation in specific cases.

[1]We acknowledge that there are several types of obligations, but this distinction is not important for the present paper.

How do norms influence the behaviour of the agent? Note that norms cannot be incorporated as a simple filter on the possible goals of an agent. For in that case the agent would always obey the norms (if feasible), whereas we want the decision to obey the norm to be a motivated 'conscious' decision. So the architecture should allow for some facility for reasoning about applying the norms and subsequent combination of the result with obligations and with the goals of the agent. The combination of norms, obligations and goals will determine the actual behaviour of the agent. Similarly, we cannot say that a norm implies an intention, but on the other hand the existence of a norm can influence the intentions of the agents. Reasons for *not* generating an intention from a norm which is applicable to the current situation include that the norm conflicts with other norms or obligations, or conflicts with existing intentions. For example, it may be a norm in an academic organisation to attend the weekly seminars, but today that may conflict with an obligation to make a telephone call at the same time as the seminar.

Taking such matters into account adds some complexity to agent reasoning. In brief, to accommodate the influences of applicable norms and obligations, we add extra steps in the basic deliberation cycle involving a notion of *deontic events*, as discussed in section 3. Of course the question arises as to how the various types of conflicts mentioned above can be identified and resolved. In this paper we do not offer a complete solution, but propose an approach based around different orderings.

Formally, we represent obligations and norms using a preference-based dyadic deontic logic, Prohairetic Deontic Logic (PDL)[20]. The standard Kripke models of PDL include a binary accessibility relation that is interpreted as a preference ordering over possible worlds. We allow multiple such preference orderings, to support different types of obligations and norms and admitting conflicting norms and obligations. For obligations, the preference ordering is related to "penalties" imposed for violation. For norms, the preference ordering is related to the "social benefit" attached to different worlds.

Obligations and norms are an important tool to "glue" autonomous agents together in a multi-agent system. Obligations restrict autonomy and norms make coordination more efficient. We are interested in describing how these external social relations of the agents influence the behaviour of the agents, specifically: how do intentions arise from norms and obligations? In the rest of this paper we will try to answer this question. In the following section we will give a formal description of norms and obligations and give a sketch of a semantics that shows links between these concepts and actions of agents. In section 3, we give an agent architecture that incorporates these social influences, including how the external influences are mapped on the in-

ternal decision mechanism of the agent. In section 3.1 we present an illustrative example in which agent behaviour is influenced by norms and obligations. In section 4 we give some conclusions and directions for future research.

2 Semantics of Norms and Obligations

We are interested in exploring the *social level* of agent behaviour, c.f. [6], and although this paper focuses on norms and obligations, many related notions such as *commitment* and *authority* are important for a full understanding of social behaviour, e.g. [2, 7, 10]. In a sense, obligations are easier to reason with than norms, because they have explicit punishments related to their violation, whereas the impact of failing to adhere to a norm can only be determined by considering a broader (and possibly longer term) impact of indirect consequences–a kind of 'second order' effect.

In order to make the distinction between norms and obligations clearer, we present an example drawn from human societies.

Suppose that in a particular organisation, all employees start work at 9 am. If this situation is a consequence of the fact that most employees first take their children to school and then come to work it is nothing more than a fact. However, it may be that it is a standard within that organisation to come to work at 9:00. Once it is seen as a standard or norm, that is in itself a reason to start work at 9:00. So, also when the children have holidays the employees will still start work at 9:00. In this case, the fact that employees start work at 9:00 am does not only have a statistical significance, it also has a social significance. In general a norm is a specific behaviour that is seen as being beneficial for the group to which a person (agent) belongs, to follow. In our example it is easier to plan meetings if it is known that employees generally start at 9:00. An important point is that, although employees know of the norm to start at 9:00, they comply to the norm of their own free will. There is no retribution if they do not comply to the norm (at least not in a direct way). Only repeated and large deviations of the norm will have consequences. For example, if many people arrive repeatedly at 8:00 am (because otherwise they cannot finish their tasks) the norm will change. If one person always arrives at 10:00 am only he/she will slowly become an "outcast" of the organisation. In practice this means that the person may have less status or his/her preferences may not be taken into account when meetings are planned.

On the other hand, if an organisation makes it a rule of conduct for its employees to start work at 9 am, then arrival at this time is an obligation. In this case there will be some direct punishment whenever this rule is violated. The punishment is usually specified together with the obligation. For example, being late more than three times (without good reason) means a cut in salary of one hour (or more) in that week.

From a utilitarian point of view one might argue that an obligation will be followed whenever the probability of getting caught while violating the obligation times the size of the punishment is higher than the expected cost of adhering to the obligation. However, this is usually not the (only) reason to adhere to an obligation! In what usually is known as a "decent" society there seems to be a norm(!) to adhere to obligations whenever possible (or applicable/reasonable). This creates another incentive to fulfil the obligation. Several cases can be distinguished now. An agent might accept that an obligation exists, but does not comply to the norm that obligations should be fulfilled. Therefore it does not fulfil the obligation for this reason. However, it might still fulfil the obligation, because the costs of violating it are too high. This distinction does not exist for norms. An agent can decide to either adopt a norm or not, but it will make the decision based solely on the benefits it sees in adopting the norm.

2.1 Formal semantics

The formal semantics of obligations (and norms) is based on Prohairetic Deontic Logic (PDL) [20]. PDL is a logic of dyadic obligation defined axiomatically in terms of a monadic modal preference logic. Only dyadic obligation is defined, i.e., all obligations are conditional, $O(p|q)$, however unconditional obligation can be represented using a tautology for the condition, $O(p) = O(p|q \vee \neg q)$. PDL allows the representation of *contrary-to-duty* obligations (obligations that hold in in sub-ideal circumstances) without contradiction, yet true deontic conflicts (conflicting obligations) imply inconsistency.

We extend PDL to allow for *multiple* modalities to denote norms and obligations from different "sources". Norms for different societies are distinguished, as are obligations to different individuals or within different organisational contexts. We take the view that obligations from the same source must be consistent, but it is allowable for obligations from two *different* sources to conflict. You can't simultaneously have two obligations to Bill: one to achieve p and the other to achieve $\neg p$. However you can have an obligation to Bill to achieve p and an obligation to Chris to achieve $\neg p$:

- $N^z(p|q)$ – it is a norm of the society or organisation z that p should be true when q is true.
- $O^z_{ab}(p|q)$ – when q is true, individual a is obliged to b that p should be true. z is the organisation/society that is responsible for enforcing the penalty

The semantics of each modality is based on a preference ordering over worlds, unique to the modality, and an equiva-

lence relation, POS, common to all modalities, that is used to interpret "possibility".

The preference ordering of norms is based on a preference of social benefit of a situation, while the preference ordering of obligation is based on the punishment when violating the obligation. For each society, x, each state, w, has a social worth, $SW(w,x)$, that defines the preference ordering for the operator N^x. In the same way for each state, w, there is a value of that world for an individual a, with respect to its relation to individual b and society x: $PW(w,a,b,x)$. This value can be seen as the cost of the punishment in case a does not fulfil its obligation towards b and defines the preference ordering for the operator O^x_{ab}.

We now follow [20] for describing a preference semantics of the conditional norms and obligations. Refer to [20] for an extensive explanation of the choice of operators, which might not always be obvious.

Start with three sets of monadic modal operators $\Box\!\!\!\!\Box$, $\Box^N_x$, and $\Box^O_{a,b,x}$. The formula $\Box\!\!\!\!\Box p$ can be read as "p is true in all possible worlds" defined in terms of the access condition, POS, which is required to satisfy the minimal constraints below. The formula $\Box^N_x p$ can be read as "p is true in all worlds that are preferred according to the norms of society x". The formula $\Box^O_{a,b,x} p$ can be read as "p is true in all worlds that are preferred according to the obligations of a towards b with respect of society x". As usual $\Diamond p \equiv \neg\Box\neg p$.

$M,w \models \Box\!\!\!\!\Box p$ iff $\forall w' \in W$ if $POS(w,w')$ then $M,w' \models p$
$M,w \models \Box^N_x p$ iff $\forall w' \in W$ if $SW(w,x) \leq SW(w',x)$, then $M,w' \models p$
$M,w \models \Box^O_{a,b,x} p$ iff $\forall w' \in W$ if $PW(w',a,b,x) \leq PW(w,a,b,x)$, then $M,w' \models p$

The $\Box^N_x$ and $\Box^O_{a,b,x}$ are S4 modalities, while the $\Box\!\!\!\!\Box$ is an S5 modality. Assume that if $SW(w,x) \leq SW(w',x)$ or $PW(w',a,b,x) \leq PW(w,a,b,x)$ then also $POS(w,w')$.

From the monadic operators $\Box^N_x$ and $\Box^O_{a,b,x}$, define binary "betterness" relations for the norms and obligations: $p \succ^N_x q$ states that "p is preferred according to the norms of society x to q". More precisely, it holds in a world w if for all possible worlds w_1 where $p \wedge \neg q$, and w_2 where $\neg p \wedge q$, w_2 is not preferred to w_1. Introduce $\succ^O_{a,b,x}$ similarly.

$$p \succ^N_x q \equiv \Box\!\!\!\!\Box((p \wedge \neg q) \rightarrow \Box^N_x \neg(q \wedge \neg p))$$

$$p \succ^O_{a,b,x} q \equiv \Box\!\!\!\!\Box((p \wedge \neg q) \rightarrow \Box^O_{a,b,x} \neg(q \wedge \neg p))$$

Also from the monadic operators, define $Id^N_x(p|q)$ and $Id^O_{a,b,x}(p|q)$. [We use the non standard notation Id rather than I to avoid later confusion with intentions.] These state that of all the worlds that are possible from the current world, (i) in all the maximally preferred (ideal) worlds where q holds, p also holds, and (ii) in all infinite chains of increasingly preferred worlds, p eventually holds:

$$Id^N_x(p|q) \equiv \Box\!\!\!\!\Box(q \rightarrow \Diamond^N_x(q \wedge \Box^N_x(q \rightarrow p)))$$
$$Id^O_{a,b,x}(p|q) \equiv \Box\!\!\!\!\Box(q \rightarrow \Diamond^O_{a,b,x}(q \wedge \Box^O_{a,b,x}(q \rightarrow p)))$$

Finally, define a *norm*, $N^x(p|q)$, or *obligation*, $O^x_{ab}(p|q)$, to be that not only is $p \wedge q$ preferred to $\neg p \wedge q$ but also the preferred (or ideal) q-worlds all satisfy p.

$$N^x(p|q) \equiv ((p \wedge q) \succ^N_x (\neg p \wedge q)) \wedge Id^N_x(p|q)$$
$$O^x_{ab}(p|q) \equiv ((p \wedge q) \succ^O_{a,b,x} (\neg p \wedge q)) \wedge Id^O_{a,b,x}(p|q)$$

3 Agent Architecture

We now explore the influences of social obligations and norms on the deliberation process of a BDI agent.

We review the abstract architecture for an isolated (non-social) BDI agent. Following [14], we can view a BDI agent as having dynamic data structures corresponding to the agent's belief, desires, and intentions, together with an event queue that keeps track of events that the agent is to respond to. Events are generated by information coming from outside the agent (external events), changes in the state of the agent, such as belief changes (internal events), and the execution of subgoals (goal events). Encoding possible ways of responding to these events is a set of plans, $\{plan(\phi,\psi,\beta),...\}$, where each plan consists of: an invocation condition, ϕ, which is the event that the plan responds to, a context condition, ψ, stating conditions under which to use the plan, and a body, β, that specifies a sequence of actions or subgoals to achieve. The main interpreter loop is essentially as follows:

BDI-interpreter
initialize-state();
repeat
    selected-events := event-selector(event-queue);
    plan-options := option-generator(selected-events);
    selected-plan-options := deliberate(plan-options);
    update-intentions(selected-plan-options);
    execute();
    get-new-external-events();
    drop-successful-attitudes();
    drop-impossible-attitudes();
end repeat

The plan-options are alternative plans to execute. In systems such as PRS and dMARS, the event-selector selects a single event from the event queue. The option generator enumerates all plans with that event as the triggering condition and with a context condition that is believed true. The deliberation to select which plan to use is based on meta-plans or hardwired strategies (for efficiency).

What impact does the existence of norms and obligations have on this model? We will concentrate on the intention generation rather than the intention execution aspects of the problem.

Firstly, we need to be able to reason about norms and obligations, so we will include explicit representations of norms and obligations. In a dynamic society, these norms and obligations may change. As well, an agent may have an incomplete or incorrect understanding of the norms and obligations that apply to itself and others. Thus norms and obligations can be considered a form of beliefs. However, in this paper we will not deal with the issues involved in communicating and updating norms and obligations, and will assume that the norms and obligations are fixed in advance.

We introduce a new sub-class of internal events, *deontic events*, corresponding to the immediate applicability of norms and obligations. For example, if the agent A has a conditional obligation to B, $O^x_{AB}(\phi|\psi)$, and the precondition ψ becomes true, then a deontic event $\mathcal{O}^x_B(\phi)$ is posted on the event queue of A. The deontic events are generated from changes in the norms, obligations, and beliefs of an agent.

To respond to these deontic events, each agent has plans whose invocation condition are deontic events such as $\mathcal{O}^x_a(\phi)$. For example, Suppose our agent desires to follow obligations (and norms) whenever they are applicable. An obvious plan would be one with invocation condition $\mathcal{O}^x_a(\phi)$ and a body to achieve ϕ.

There are various reasons why an agent might not automatically follow a norm or obligation:

- the precondition is not satisfied, so the norm is not applicable;
- the precondition is satisfied, so the norm is applicable, but the agent has not adopted the norm (not dealt with in this paper);
- the norm is applicable and adopted but:
  - conflicts with other norms or obligations; or
  - conflicts with existing intentions;
- the norm is applicable and adopted, but the agent reasons that the norm does not achieve its original intention and in fact makes matters worse in this case (not dealt with in this paper).

Note that the existence of a norm or obligation may not just lead to adopting a plan to satisfy that norm or obligation. For example, an unscrupulous agent might respond to a particularly burdensome obligation by adopting a plan to *evade* rather than *meet* the obligation, say by leaving the country.

Some norms and obligations deal directly with the intentions of the agent. For example, suppose Al has an obligation to perform a task for Bob (with respect to organisation z) – Al may or may not intend to meet his obligation, but he also has a norm that he should tell Bob if he does not intend to meet his obligation. This norm deals not with external properties of the world, but with Al's internal state. This means that the agent needs to be able to explicitly represent its intentions within its representation of norms and obligations, and that such intentions are part of any language used for communication of norms and obligations between agents.

Let us represent the fact that the agent intends to satisfy an obligation, $\mathcal{O}^z_B(\phi)$, by $I(\phi)$. If τ_B represents telling Bob about meeting the obligation, then the introspective norm would be written as $N^z(\tau_B|\mathcal{O}^z_B(\phi) \wedge \neg I(\phi))$.

The existence of these introspective norms and obligations leads to further complications. Consider Al's obligations above, until he actually commits to not meeting his obligation to Bob, the need to tell Bob does not exist, yet the *potential* for it may have a significant impact on his decision on whether to do the task for Bob. For example, imagine that the task itself is trivial (i.e., the direct consequences of not doing the task are small), but the social consequences of not informing Bob are very high (e.g., Al is perceived as unreliable). If for some reason Al is not able to inform Bob, then the existence of the norm should affect his decision on whether to intend to fulfil his obligation.

Consider an iterative approach, where Al decides not to do the task in one cycle of the interpreter loop and then only when the norm actually becomes applicable on the next cycle does he consider the consequences. Al must either live with the consequences of the decision or try to undo his commitment – neither being pleasant options!

Instead, we introduce *potential deontic events*. These are events that may also exist, depending upon what plan-options are chosen in the deliberation step. The option-generator thus needs to take some set of events from the event queue to react to and then add in all the potential deontic events that exist, by virtue of these introspective norms and obligations:

Thus we have an additional step in the interpreter loop, where the selected events are augmented with potential deontic events generated by repeatedly applying the introspective norms and obligations. We will handle the constraints between obligations in the option-generator step, rather than the deliberation step. Thus the plan-options in the modified interpreter loop are possible sets of plans to intend simultaneously. The deliberation step then selects between these sets of plans on the basis of the preferences.

modified BDI-interpreter
initialize-state();
repeat
selected-events := event-selector(event-queue);
augmented-events :=
potential-event-closure(selected-events)
plan-options :=
option-generator(augmented-events);
selected-plan-options := deliberate(plan-options);
update-intentions(selected-plan-options);
execute();
get-new-external-events();
drop-successful-attitudes();
drop-impossible-attitudes();
end repeat

In the generation of intentions in the modified interpreter loop, there are two places where choices are made: event-selector and deliberate. The "event-selector" choice selects some subset of "most important" events, and the "deliberate" choice determines which of the alternative courses of action should be used to respond to these events.

We would expect a socially responsible agent to make decisions consistent with the externally defined orderings over the norm and obligations, e.g. [4].

Let us consider the example of an information-gathering agent, Al, servicing requests from other agents in an organisation, z. Al has agreed to provide services to Bob, his boss, and Chris, a co-worker. The contractual obligation is expressed by two obligation expressions, Al is obliged to Bob to achieve ϕ_B when ψ_B and Al is obliged to Chris to achieve ϕ_C when ψ_C: $O^z_{AB}(\phi_B|\psi_B)$ and $O^z_{AC}(\phi_C|\psi_C)$. Given Bob is Al's boss, and Chris is Al's co-worker, there is a ordering over the obligations, with Al's obligations to his boss being more important than those to his co-worker, $O^z_{AB}(\phi_B|\psi_B) > O^z_{AC}(\phi_C|\psi_C)$. Suppose both of these obligations became applicable at the same time and they turned out to be incompatible. For Al to be socially responsible, it would have to fulfil its obligation to Bob at the expense of the obligation to Chris (assuming no other influences).

Note that there may be other factors that allow Al to override this ordering over its obligations. For example, it may be a norm of an organisation that you do not overburden colleagues undergoing severe personal stress. This may be partly interpreted as a norm that says you meet your obligations to such individuals: $N^z(\phi|\mathcal{O}^z_x(\phi) \wedge stressed(x))$. If Chris is under personal stress, then it is possible for Al to fulfil its obligation to Chris at the expense of fulfilling its obligation to Bob, and still be a socially responsible agent.

Within the constraints imposed on a socially responsible agent by the ordering over norms and obligations, there is flexibility, since the ordering is only partial. There may be no objective comparison between certain norms and obligations, especially those referring to different societies or organisations. For example, if Al had an obligation to Chris by virtue of something other than the relationship through organisation z, say a family relationship, then it might not be so easy to compare the obligations. This means that different agents are allowed to make different decisions, based on a "subjective" preference between otherwise unordered norms and obligations. This can be perceived as part of the different "personalities" of different agents.

The deliberation step needs to choose between the various plan-options. The orderings over norms and obligations described in Section 2 can be used to assist with the deliberation. However, the different orderings for different societies or organisations are not necessarily comparable in any objective way. An agent can "subjectively" reduce the different orderings must to a single (possibly partial) order over all the norms and obligations. Different agents may value the same norms and obligations differently and these differences can be viewed as "personality" traits of the agents.

For example, c.f. [9], explicit consideration of interactions between an agent's norm following behaviour and its approach to managing goal and intention conflicts, allows the characterisation of different 'personalities' of agents within this single framework. For example, a "legalistic" agent might rate all obligations more highly than its norm, whereas a "social conformist" agent might rate norms more highly than obligations.

3.1 Discussion

Let us further consider the above example of Al, Bob, and Chris.

Within agent Al there are representations of the two obligation expressions that can lead to the generation of two deontic events, $\mathcal{O}^z_B(\phi_B)$ and $\mathcal{O}^z_C(\phi_C)$. To respond to these events, the agent has plans, $plan(\mathcal{O}^z_B(\phi_B), \psi'_B, \beta_B)$ and $plan(\mathcal{O}^z_C(\phi_C), \psi'_C, \beta_C)$.

Suppose a condition ψ occurs where $\psi \rightarrow \psi_B \wedge \psi_C \wedge \psi'_B \wedge \psi'_C$ and where ψ makes it impossible to achieve both ψ_B and ψ_C, $\psi \rightarrow \neg(\phi_B \wedge \phi_C)$. This condition causes the events $\mathcal{O}^z_B(\phi_B)$ and $\mathcal{O}^z_C(\phi_C)$ to be added to the event-queue. Let us suppose that these events are the only events now in the event queue.

The first step inside the loop is to select events to deal with. We can either choose one event, and deal with the other and the incompatibility of the two responses in the next cycle, or we can choose to look at both events.

If we choose only one event, the existence of the ordering over the two obligations (and an assumption that Al is socially responsible) requires Al to consider the more important obligation, the one to Bob, first. We would then deal with the other event and the incompatibility of the responding to both in the next cycle.

Let us assume we choose both events. There are no introspective norms and obligations to worry about, so no potential deontic events need be added during the potential-event-closure step.

The option-generator step takes into consideration the incompatibility of the two obligations, and generates the following three plan-options (using the body of the plan to stand for the plan itself): $\{\beta_B\}$, $\{\beta_C\}$, and $\{\}$ (i.e., doing nothing is also an option). In the deliberation step, the ordering on the obligations is taken into consideration, and $\{\beta_B\}$ is selected.

Now let us consider what happens if we add in an introspective norm that states that if you are obliged to do a task for someone, x, and don't intend it, then you should ensure that they are told, τ_x: $N^z(\tau_x|\mathcal{O}^z_x(\phi) \wedge \neg I(\phi))$. We also need a plan $plan(\tau_x, true, doTell_x)$.

The event-selector generates the same set of events as before, $\{\mathcal{O}^z_B(\phi_B), \mathcal{O}^z_C(\phi_C)\}$. Now at the potential-event-closure step, we need to add in the potential deontic events $\mathcal{N}^z(\tau_B)$ and $\mathcal{N}^z(\tau_C)$.

The option generator now generates the following set of options. Note that plans reacting to the potential deontic events are only considered when the norm is relevant:

- $\{\beta_B, doTell_C\}$ – achieve ϕ_B and inform Chris
- $\{\beta_B\}$ – achieve ϕ_B and don't inform Chris
- $\{\beta_C, doTell_B\}$ – achieve ϕ_C and inform Bob
- $\{\beta_C\}$ – achieve ϕ_C and don't inform Bob
- $\{doTell_B, doTell_C\}$ – achieve nothing and tell Bob and Chris
- $\{doTell_B\}$ – achieve nothing and tell Bob
- $\{doTell_C\}$ – achieve nothing and tell Chris
- $\{\}$ – achieve nothing and tell nobody

In the absence of any other influences, a socially responsible agent would select $\{\beta_B, doTell_C\}$ at the deliberation step. In the presence of the belief $stressed(C)$ and a norm $N^z(\phi|\mathcal{O}^z_x(\phi) \wedge stressed(x))$ that is not objectively comparable to the other norm and obligations, it is possible for Al to select either $\{\beta_B, doTell_C\}$ or $\{\beta_C, doTell_B\}$, depending upon Al's subjective preferences.

4 Concluding remarks

For simplicity, in this paper we have only dealt with the external influence on an agent - the social norms and obligations imposed from the outside. Also relevant to the behaviour of autonomous agents are the internal driving forces: the long-term *intrinsic goals* of the agent that represent the agent's "personal desires" independent of what society says the agent ought to do. The goals of the agent may be at odds with the agent's norms and obligations (indeed, the explicit punishment for violation of obligations can be seen as a way of countering the opposing influence of the agent's personal desires).

The intrinsic goals of an agent can be incorporated into the framework described above in a manner similar to norms and obligations.

- A goal $G(\phi|\psi)$ represents a personal desire that ϕ be true whenever ψ holds (as in the case of norms and obligations, unconditional goals can be represented using a tautology for the condition).
- Within the interpreter loop we introduce goal events $G(\phi)$ in to the event queue whenever the precondition ψ of a goal $G(\phi|\psi)$ holds.
- When making choices within the interpreter loop, the intrinsic goals have to be balanced against the norms and obligations, just as the norms and obligations need to be balanced against one another.
- Semantically, goals are treated as another form of modal operator with new preference relations. Whereas the preference relations for norms reflect "social benefit" and the preference relations for obligations reflect penalties for violation, the preference relations for goals reflect a measure of personal utility for the agent.

In this context we can consider new types of "personality" based on the relative importance attached to goals, norms, and obligations: "selfish" – intrinsic goals are rated more highly than norms or obligations; "survivalist" – goals related to the continued existence of the agent are rated more highly than any other influences; "indolent" – the agent has a goal to expend as little effort as possible, even in the face of quite strong norms and obligations.

Our architecture is different from that of agents that make decisions among different behavioural alternatives on the basis of utility and probability , e.g. [11, 15]. Such agents are autonomous since they are self- interested. However, this type of agent usually has two problems. First they have a fixed social attitude, i.e. they are either selfish or altruistic or something in between. They cannot change from being altruistic into selfish after being cheated by an agent or differentiate their behaviour with respect to different agents. Another problem is that the influence of the norms is fixed by a static utility function. The consequence is that the agent cannot (easily) take a norm into account in different ways according to the circumstances. E.g. a norm

not to delete files might be easily violated when the file is known to contain a virus.

A strong motivation for our work is to find a framework for building agents that can exhibit a wide range of collaborative behaviours [7], which may include tightly coordinated teamwork, e.g. [18], or could just involve the exploitation of a variety of more flexible social behaviours [2]. In this context, obligations include support for efficient execution of joint plans [8, 18], and for characterising certain forms of social commitment [4]. They provide additional stability to joint plans in that once the parties have agreed on a plan, the adoption of obligations makes it less likely that one member of the team will suddenly decide not to play its part – there is an additional penalty on the team member beyond just the failure of the original goal, which the member may no longer see as important. Equally important in this larger context is the consideration of obligations that arise from an agent's position in society, in particular taking into account various forms of power or dependence relationships and ability to delegate. Such arrangements are often used to describe *organisational structures*, e.g. [19]. In future work we plan to deal carefully with such issues, as has begun to be illustrated in the example above (Section 3.1).

The framework presented in this paper is complementary to that of [3], where an approach to modelling normative reasoning is presented within the DESIRE framework. In this paper, we focus, in effect, on the process of generating (candidate) intentions, in the context of extending a standard BDI architecture to accommodate norms and also obligations. Future work will address more fully the development of appropriate semantics, leading (we hope) to a theory which will support reasoning about the respective preference relationships, and characterisations of properties of the orderings which should be respected by the various selection functions which are embedded in the extended BDI interpreter. In addition to this theoretical perspective, we will seek to contribute in the area of social simulations, e.g. [12], exploring, for example, issues to do with the interactions between agents of different personalities, and societies of different normative structures.

Acknowledgements This work was supported by a grant from the Australian Research Council.

References

[1] Barbuceanu, M. Coordinating with Obligations. In *Autonomous Agents 98*, Minneapolis, 1998, ACM Press, pp 62-69.

[2] Castelfranchi, C. Modeling Social Action for AI Agents *Artificial Intelligence* 103(1998)157-182.

[3] C. Castelfranchi, F. Dignum, C. Jonker and J. Treur. Deliberate Normative Agents: Principles and Architectures, In N. Jennings and Y. Lesperance (eds.) *Proceedings of ATAL-99*, Orlando, 1999, pages 206- 220.

[4] L. Cavedon and L. Sonenberg On social commitments, roles and preferred goals, In *Proceedings of the 1998 International Conference on Multi-Agent Systems*, ICMAS98, Paris, July 1998, (ed) Y Demazeau, pp 80-87.

[5] F. Dignum. Information Management at a bank using agents: theory and practice, In *Applied Artificial Intelligence*, forthcoming.

[6] F. Dignum and B. van Linder. Modelling social agents: Communication as actions. In M. Wooldridge J. Muller and N. Jennings, editors, *Intelligent Agents III (LNAI-1193)*, pages 205–218. Springer-Verlag, 1997.

[7] B. Grosz Collaborative Systems *AI Magazine* 17(1996)67-85.

[8] B. Grosz and S. Kraus. Collaborative Plans for Complex Group Action, in *Artificial Intelligence* 86(1996)269-357.

[9] L. Hogg and N. Jennings. Variable Socialability in Agent-based decision making, In N. Jennings and Y. Lesperance (eds.) *Proceedings of ATAL-99*, Orlando, 1999, pages 276-290.

[10] N. Jennings. Commitments and Conventions: The foundation of coordination in Multi-Agent systems. *Knowledge Engineering Review*, vol. 8(3), pages 223-250, 1993.

[11] Jennings. N.R. and Campos, J.R. Towards a Social Level Characterisation of Socially Responsible Agents. *IEEE Proc. on Software Engineering*, vol.144, 1, pp.11-25, 1997.

[12] Journal of Artificial Societies and Social Simulation, `http://www.soc.surrey.ac.uk/JASSS/JASSS.html`.

[13] Kinny M., Ljungberg, M, Rao, A, Sonenberg, E, Tidhar, G, and Werner, E. Planned Team Activity. In *Artificial Social Systems*, Springer LNCS 830, pages 227-256 (Eds) C Castelfranchi and E Werner , 1994

[14] Rao, A.S., and Georgeff, M.P. BDI Agents: From Theory to Practice. *Proceedings of the First International Conference on Multi-Agent Systems, ICMAS 95*, San Francisco, 1995

[15] Rosenschein, J., and Zlotkin, G. Rules of Encounter. MIT Press, Cambrige, USA, 1994.

[16] Shoham, Y. and Tennenholtz, M. On social laws for artificial agent societies: off-line design. *Artificial Intelligence* 73(1995)231-252.

[17] M. Singh Multiagent Systems as Spheres of Commitment. In *International Conference on Multiagent Systems (ICMAS) Workshop on Norms, Obligations, and Conventions*, Kyoto, Japan, December 1996.

[18] Tambe, M. Agent architectures for flexible, practical teamwork. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-97)* , 1997.

[19] Tidhar, G. and Sonenberg, E.A., Organized Distributed Systems. Submitted for publication, March 2000.

[20] L. van der Torre and Y.-H. Tan. Contrary-To-Duty Reasoning with Preference-based Dyadic Obligations. In *Annals of Mathematics and AI*, submitted.