A Bayesian Approach to Norm Identification*

(Extended Abstract)

Stephen Cranefield and Tony Savarimuthu
University of Otago
Dunedin, New Zealand
{stephen.cranefield, tony.savarimuthu}@otago.ac.nz

Felipe Meneguzzi
Pontifical Catholic University of Rio Grande do Sul
Porto Alegre, Brazil
felipe.meneguzzi@pucrs.br

Nir Oren
University of Aberdeen
Aberdeen, UK
n.oren@abdn.ac.uk

ABSTRACT

When entering a system, an agent should be aware of the obligations and prohibitions (collectively *norms*) that will affect it. Several solutions to this *norm identification* problem have been proposed, which make use of observations of either other's norm compliant, or norm violating, behaviour. These solutions fail in situations where norms are typically violated, or complied with, respectively. In this paper we propose a Bayesian approach to norm identification which operates by learning from both norm compliant and norm violating behaviour. By utilising both types of behaviour, our work not only overcomes a major limitation of existing approaches, but also yields improved performance over the state-of-the-art. We evaluate its effectiveness empirically, showing, under certain conditions, high accuracy scores.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*intelligent agents, multiagent systems*

General Terms

Algorithms, Design, Theory

Keywords

Norm recognition; Norm identification; Bayesian reasoning

1. INTRODUCTION

Within the multi-agent systems community, norms have been viewed as a means to provide declarative control over agent behaviour while preserving their autonomy. Norms are instantiated as obligations, prohibitions and permissions under specific circumstances. In turn, these act as soft constraints, specifying the behaviour expected of the agent. However, agents can *violate* these constraints for a variety of reasons, including the pursuit of an important goal; irrational behaviour; or maliciousness. When violating a norm, an agent typically has a sanction imposed upon it by some entity within the system. A large body of work exists investigating how norms can be used to constrain behaviour of software agents [1]. Such work typically focuses on formal semantics; practical reasoning; and norm emergence.

Much less attention has been paid to the problem of *norm identification*, which considers how an agent can identify norms already present in a system. While existing work usually assumes that agents are aware of all norms that might affect them, this assumption is unrealistic, particularly in open multi-agent systems, where new agents can join or leave the system at any time, and where factors such as limited bandwidth could hinder the transfer of norms. Other situations where norm identification becomes important include systems where norms are implicit rather than formally specified; where agents are malicious (and could therefore lie regarding the existence of a norm); and where there is no shared ontology to facilitate communication between agents. In such situations, agents must be able to detect, learn or identify norms so as to act appropriately, and reduce the risk of being sanctioned.

Savarimuthu et al. [4] proposed a typology of norm identification methods. Among these, *observation-based* techniques are perhaps the most popular. Here, agents observe the behaviour of others to infer a system's norms. Savarimuthu et al. [4, 3] proposed one such technique based on the detection of *violation signals*, representing optional punishments or sanctions imposed by agents after observing another violating a norm. By learning the situations in which these violation signals arise, an agent can infer a signal's triggering norm. While effective in the presence of norm violations, such an approach is difficult to apply in systems where agents (largely) comply with norms. In response, Oren and Meneguzzi [2] proposed an alternative norm inference mechanism. In their approach, an agent infers the goals pursued by others by applying plan recognition to observed action sequences. By considering the states and actions always avoided or achieved in pursuit of goals, prohibitions and obligations are identified. However, their basic approach does not function well in the presence of norm violations, and an extension that can cater for such violations requires prohibitively large amounts of memory to function.

Both approaches discussed thus far, as well as others described in the literature, function well in extreme cases when

*See `http://hdl.handle.net/10523/5476` for the full version of this paper.

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*

norms are (nearly always) violated or complied with. However, they perform poorly in situations where both norm compliance and violation regularly occur. Our core contribution is to suggest a new approach to norm identification which operates well in such situations. Our approach makes use of both a violation signal and plan recognition, together with Bayesian reasoning to associate a likelihood ratio with an obligation or prohibition. An agent can then use these ratios to pick and comply with the most likely norms. We show that through the use of our techniques, an agent can act in a norm-compliant manner after relatively few observations of the behaviour of others.

2. MODEL OVERVIEW

We consider the identification of norms governing transitions of agents through a graph encoding a state space. Transitions between states are the result of an agent executing a plan with the goal of transitioning from a start node to a destination node. Plans are thus sequences of nodes. We make no assumptions about the source of plans: they may be generated dynamically given a goal and a set of actions, or may come from a plan library, such as a BDI agent program. Our norm identification mechanism is based on the assumption that the observed agents' plan libraries (or available actions and planning mechanism) are known to the observing agent. This would be the case if all agents share the same plan library, at least at some level of abstraction, but can also be seen as a hypothesis made by the observer to gain some traction on the norm identification problem.

Our norm hypothesis space is defined by a subset of linear temporal logic comprising three norm types and their negations. Informally, these are: *i)* $eventually(n)$ / $never(n)$: unconditional norms constraining a plan execution to include or exclude node n; *ii)* $next(cn, n)$ / $not\_next(cn, n)$: conditional norms stating that if 'context node' cn is reached, it must or (respectively) must not be followed by node n (we only consider norms of these types when an edge from cn to n exists in the graph); and *iii)* $eventually(cn, n)$ / $never(cn, n)$: also conditional norms, expressing that *beginning from the node after the context node*, node n will be eventually or (respectively) never reached.

We assume that agents can observe paths traversed by other agents in the graph. In addition, in line with the work of Savarimuthu et al. [4, 3], we assume that agents can detect *signalling actions* that indicate sanctioning of the observed agent. These signals may indicate a sanction applied after a norm has been breached, or may be non-normative signals emitted by agents due to their own values or personal norms being breached (we refer to these, respectively, as sanction and punishment signals, and collectively as violation signals). We model the latter case by assuming there is a small population-wide probability of a non-normative punishment signal being observed after any step of an observed path. Furthermore, we associated fixed probabilities with the likelihood of norm violations being observed, and of observed violations being sanctioned.

Given an observed trace, for each possible norm we compute the conditional likelihood of (a) the trace and observed violation signals, and (b) the trace given the plan library. Bayes' rule is then used to update the odds of each norm compared to a null hypothesis that no norm exists.

3. RESULTS SUMMARY

In order to evaluate our work, we utilised our approach as a decision mechanism for action: an agent begins by observing others, and then acts based on the norms it has learned. We then computed precision and recall scores for the norms it believed existed within the system. Our experiments show that our mechanism, when selecting the top ten norms ranked by odds, results in significant precision and recall scores. Even in a graph with almost 2000 possible norms, our techniques enable an agent to generate norm-compliant plan executions most of the time without any prior knowledge of the active norms within a system, achieving an F_1 score as high as 0.95. Finally, our experiments suggest a significant increase in the approach's effectiveness when a violation signal can be used, with a significantly better F_1 score in such cases.

4. CONCLUSIONS AND FUTURE WORK

Previous work on observation-based identification of norms has produced approaches that use evidence either assuming that agents are fully compliant with norms or that agents violate norms and that such violations produce an observable signal. This is a serious limitation as such techniques do not utilise all information available from the behaviour of an agent that can both comply with, and violate, norms in different situations. This work addresses this gap by considering both types of evidence and employing Bayes' rule to compute the odds of each possible norm compared to the absence of any norms. It then uses the resulting odds to identify norms via a ranking mechanism. We empirically demonstrated the effectiveness of our approach on a range of scenarios of varying complexity and size, and examined the impact of violation signals as a source of information. We found that at low levels of violation our approach can generate norm-compliant behaviour at least 70% of the time in the presence of a large number of norms, and up to 99% of the time for societies with a small number of norms. Importantly, we show that when agents do violate norms, and when such violations are observable, we substantially improve recall for detecting such norms.

Acknowledgments. Felipe thanks CNPq for support within process numbers 306864/2013-4 under the PQ fellowship and 482156/2013-9 under the Universal project programs.

REFERENCES

[1] G. Andrighetto, G. Governatori, P. Noriega, and L. W. N. van der Torre, editors. *Normative Multi-Agent Systems*, volume 4 of *Dagstuhl Follow-Ups*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2013.

[2] N. Oren and F. Meneguzzi. Norm identification through plan recognition. In *Proceedings of the workshop on Coordination, Organization, Institutions and Norms in Agent Systems (COIN 2013@AAMAS)*, 2013.

[3] B. T. R. Savarimuthu, S. Cranefield, M. A. Purvis, and M. K. Purvis. Obligation norm identification in agent societies. *Journal of Artificial Societies and Social Simulation*, 13(4):3, 2010.

[4] B. T. R. Savarimuthu, S. Cranefield, M. A. Purvis, and M. K. Purvis. Identifying prohibition norms in agent societies. *Artificial intelligence and law*, 21(1):1–46, 2013.