CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Dagstuhl Research Online Publication Server

Norms and accountability in multi-agent societies (extended abstract)

Rodger Kibble[1]

Goldsmiths, University of London, Department of Computing
Lewisham Way, London SE14 6NW, UK
r.kibble@gold.ac.uk

Abstract. It is argued that *norms* are best understood as classes of constraints on practical reasoning, which an agent may consult either to select appropriate goals or commitments according to the circumstances, or to construct a discursive justification for a course of action after the event. We also discuss the question of how norm-conformance can be enforced in an open agent society, arguing that some form of peer pressure is needed in open agent societies lacking universally-recognised rules or any accepted authority structure. The paper includes formal specifications of some data structures that may be employed in reasoning about normative agents.

Keywords. Norms, agents, social commitments, reasoning

1 Introduction

Researchers in multi-agent systems have often looked to analytic philosophy for suitable concepts and theoretical frameworks; indeed it could be said that philosophers since Aristotle have been engaged in writing specifications for rational agents. Some influential approaches have included Bratman's work on practical reasoning [1] and Austin and Searle's speech act theories [2,3]. Kibble [4] offered a critique of approaches to ACL semantics based on Speech Act theory such as FIPA's ACL [5] and outlined an alternative commitment-based approach drawing on more recent philosophical studies by Robert Brandom [6,7] and Joseph Heath [8]. Brandom's work presents an *inferentialist* account of theoretical and practical reasoning and communication, arguing that mentalist notions such as belief can be dispensed with in favour of more precise notions of observable practical and propositional *commitments* (though it turns out that this term does not seem to have a uniform interpretation among analysts; see [9] for discussion). Heath works at the frontiers of social theory and analytic philosophy, developing an account of the interaction of instrumental rational choice and social norms via a critical engagement with the work of Habermas [10,11]. Kibble [4] drew on this work with the aim of extending and elaborating the non-mentalist *social commitment* model of agency of [12]. Kibble proposed

Dagstuhl Seminar Proceedings 07122
Normative Multi-agent Systems
http://drops.dagstuhl.de/opus/volltexte/2007/916

an account of agent communication as *norm-governed action* for an agent to produce a dialogue act is to take on certain *commitments*, such as to defend the content of an assertion if challenged, and other agents are bound to concede that the agent is *entitled* to the propositional content of dialogue acts if those commitments are fulfilled. The present paper continues to build on this work, attempting to reconstruct and/or extend some of Brandom's and Heath's proposals concerning norms, sanctions and commitments in a form that can be applied to interactions between software agents. The paper is structured as follows:

- Section 2 considers how norms can be maintained by *peer pressure* rather than authoritarian structures of command and control, adopting elements of Heath's account of social norms;
- Section 3 argues that social norms can be represented as *constraints* on practical reasoning, rather than more primitive entities such as *goals* or *commitments*;
- Section 4 specifies some data structures to support reasoning about normative agents, adopting the notation of d'Inverno and Luck's SMART framework [13].

2 Norms, commitments and sanctions

As with multi-agent research in general, the study of normative agents suffers from inconsistent use of terminology and lack of consensus on the meaning of some fundamental terms: what exactly are *norms*? Assuming agreement can be reached on some working definitions, one of the questions that then has to be addressed is: why do (or should) agents conform to norms? There have been suggestions that failure to honour normative commitments should be subject to *sanctions*, but there have been few concrete proposals as to what form these sanctions might take or who is to be responsible for administering them.

At a certain level of abstraction we can consider norms as solutions to coordination games [14] that cannot be accounted for simply in terms of maximising utility. Classical game theory models agent interactions as a matrix of the payoffs for each agent according to the actions independently chosen by all players (*strategies*), and assumes both that the potential payoffs are known in advance to all players and that they will converge on a "Nash equilibrium" such that neither player could increase their payoff by changing strategy [15]. Non-trivial interactions tend to have many such equilibria however, yet the smooth functioning of society relies on some particular solution being commonly accepted, and thus on agents having some mechanism at their disposal for persuading or encouraging other agents to stick to the shared rules.

Moving towards the concrete: Brandom [7, 84ff] discusses three classes of norm involved in practical reasoning: the *prudential* or instrumental, *institutional* and *unconditional*, illustrated in the following examples:

α. Only opening my umbrella will keep me dry, so I will open my umbrella.

β. I am a bank employee going to work, so I shall wear a necktie.
γ. Repeating the gossip would harm someone, to no purpose, so I shall not repeat the gossip.

How would violations of these norms be sanctioned, if at all? In example (α), the "prudential" norm is a personal *preference* to stay dry rather than a social norm, so the only likely "punisher" is Nature rather than any human agent - unless perhaps the speaker is en route for some event where it would be inappropriate to turn up with wet hair and soaked clothes. The institutional norm, of wearing a necktie and being otherwise soberly dressed while working at a bank, is most likely reinforced by the threat of disciplinary action or even dismissal; minor violations might only be subject to disapproving comments from co-workers or clients. Finally, what Brandom refers to as "unconditional" rules about avoiding unnecessary harm and generally behaving in an ethical and considerate manner are not subject to institutional sanction: violations might be punished by the "voice of conscience" or if they became widely known, by expressions of reproach from friends, family etc. As Heath [8, p. 154] notes, sanctions against violations of social codes tend to be *symbolic*, intended to produce feelings of shame and regret rather than to physically harm or hinder the offender, and to articulate and re-affirm the norm.

Carrying across these distinctions into normative multi-agent systems will not be straightforward: for one thing we can safely assume that software agents are not subject to feelings of shame. Lopez y Lopez et al [16] for example use "norm" as an umbrella term encompassing "obligations, prohibitions, commitments and social codes", which would appear to fall under Brandom's headings of "institutional" and "unconditional". My approach proceeds from rather different assumptions:

Norms vs goals: The core of the definition of norms offered by [16] is a set of *normative goals* which specify "something that ought to be done". I will argue in the next section for a clear distinction between goals and norms, the latter being concerned with how goals are to be achieved and how the actions taken to achieve them can be justified.

Institutional vs bottom-up norms: The emphasis in [16] is on institutional norms, namely obligations and permissions, which are taken by [13] to be the only species of norm whose violation is punishable. Social commitments are stipulated to have rewards for compliance but no punishments for violation, while social codes have neither. Of the four different categories of artificial society described by [17], institutional norms are appropriate for closed, semi-closed and semi-open societies, where it is feasible to have commonly accepted rules and "enforcer" agents whose authority is universally recognised. With the growing potential for agent applications in open environments such as the Semantic Web, I suggest that this approach needs to be supplemented by considering whether and how norms can be sustained "horizontally" without assuming the existence of legislators, enforcers and so on.

A particular issue for MAS is indeed how norms can be enforced in open environments where norm-conformant agents interact with instrumentally rational

agents. The main lacuna is probably in the area of *sanctions*: for an agent to be socially committed entails that failure to redeem the commitment will be subject to sanction, but the literature contains few concrete proposals on the precise nature of the appropriate penalties (though Walton and Krabbe [18, pp. 20, 184] offer some tentative suggestions). Brandom proposes sanctions for nonfulfilment of a commitment [6, p. 163] though apparently not for arbitrarily withdrawing commitments [7, p. 93], and no specific sanctions are specified for failure to honour propositional commitments. A more recent proposal [19] defines violation criteria for specified types of commitment and assigns fixed numerical penalties for violations. However, the authors are silent on how these numbers might translate into effective punishments that could hinder the offending agents, and on what protocols or structures of authority could be involved in the application of sanctions.

The approach taken in this paper is influenced by Heath's discussion of social norms [8, pp. 150-161], which itself draws on the work of Durkheim and Talcott Parsons (see Heath op. cit. for references). The key ideas are:

- sanctions serve to penalise *deviance* in the sense of prioritising instrumental considerations over norms;
- norm-conformant agents are characterised not only by being disposed to follow norms themselves but by "*the disposition to punish those who do not*" [8, p. 155, emphasis in original];
- agents which are not norm-conformant by design will thus have instrumental reasons to follow norms;
- prior to being sanctioned, agents may receive the opportunity to "give an account" of their reasons for action, as it may not always be evident whether an aberrant action results from deliberate deviance, *dissent* (adherence to a different set of norms from the majority) or misunderstanding (op. cit., p. 160).

Kibble [4] proposed that agents have the following options for penalising breaches of communicative norms:

- *ostracism*: the offending agent is notified that its messages will not be accepted for a specified time period, or until it performs the requested justificatory speech act;
- *blacklisting*: the complaining agent may broadcast details of the offence to trusted agents, which may decide to implement sanctions themselves.

These penalties could in principle be generalised to other instances of norm violations. For example in an e-commerce environment, temporary exclusion from the market-place would be a highly effective sanction against dubious business practices. Furthermore, if the penalties are imposed for a fixed time period, the duration of the time period could be determined according to the numerical calculations described by [19]. For an agent to choose to conform to a norm assumes either that they are designed with the capacity to recognise and reason about normative behaviour, or that their human principals may realise that something is going wrong and decide to re-engineer or replace their agent software.

3 Norms as constraints on reasoning

The principal claim I want to argue for in this section is that a norm is not simply a goal or commitment (including negative goals, i.e. prohibitions) but a set of criteria to enable agents to select an appropriate goal or adopt an appropriate commitment according to the circumstances. To take a fairly stark example, most religious and ethical systems include a precept against killing, yet also tolerate the taking of life in certain defined situations: self-defence, as a soldier in a "just war", as a policeman dealing with a life-threatening situation and so on. So the applicable norm in such systems is not simply a prohibition, *Do not kill*, it is a class of licit inference patterns leading to a conclusion *Do not kill* or *You may kill* according to the circumstances[1].

Another way of looking at things is to adopt a *discursive* account of goals and norms: instead of considering their role in *determining* an agent's actions, we may consider how they can be invoked after the event to *account* for the actions. From this point of view we can make quite a clean separation:

- *explanations* of an action or course of actions will make reference to the agent's *goals*, perhaps supplemented by a sequence of means-end reasoning.
- *justifications* of an action additionally need to refer to *norms*: goals alone cannot justify an action, since the legitimacy of the goals themselves as well as the means employed to achieve them may be at issue.

For example, there is an apocryphal tale of a career criminal who was asked why he kept robbing banks and replied, "That's where the money is". This may count as an explanation of goal-directed action, but not as a justification. If he had said something like, "The banks destroyed my livelihood by foreclosing the mortgage on the family farm", this could be understood as an appeal to an intelligible normative framework.

We could also express this distinction by saying that goals give rise to *commitments* to actions, while norms give rise to (claimed) *entitlements* to those commitments. In fact an underlying theme of this section is the relation between norms and *responsibility*, in the sense of the following statements:

> judgement and action ... are in a distinctive sense what we are *responsible* for. They express *commitments* of ours ... [7, p. 80]

> Accountability ... captures two related aspects of the structure of norm-governed action, namely, that agents can be called upon to justify their actions vis-a-vis the relevant norms... and that they can be sanctioned for failure to comply with the prevailing normative expectations... [8, pp. 151-2]

[1] Of course, there are certain communities such as Quakers or Jains for whom the inference would be somewhat vacuous, in that the conclusion would invariably forbid killing.

The previous section discussed the second sense of accountability; here we are concerned with the first. Before proceeding further I wish to depart from Brandom's terminology in one respect. He uses the term "prudential norm" where I prefer to speak of instrumental preferences, reserving the term norm for *social* norms, where other agents' expected behaviour plays a part in deciding whether or not to conform. With reference to example α above, the speaker doubtless prefers to avoid getting wet whether or not anyone else will ever know about it. Having established this distinction, I propose that an action can be considered as norm-conformant if an agent called to account for the action is barred from offering a purely instrumental explanation. This ties in with the observation in section 2 that a norm is a solution to a coordination game which cannot be specified purely in terms of maximising payoffs.

The notion of accountability establishes a link between action and communication: various researchers in MAS have followed [20,18] in treating agent communications as actions that express or give rise to *commitments*: an agent can be said to be privately committed to the truth of a proposition, or publicly committed to producing an argument supporting the proposition if challenged. (A strictly non-mentalist account would only admit the second of these senses.) Likewise, we can say that an agent who has adopted a goal is privately committed to a course of action, and executing the action creates a public commitment on the agent to justify it if challenged, or to demonstrate *entitlement* to a set of goals and actions.

Brandom [7] stresses the inescapably non-monotonic nature of practical reasoning: in examples α - γ, the conclusion could be invalidated by an additional premise. For instance if we accept β as a good inference, the following variant β' may still be classed as bad:

> β' I am a bank employee going to work, and today is Dress-down Friday, so I shall wear a necktie.

The implication is that in the above examples of practical reasoning, the particular norm being invoked cannot simply be filled in as a missing premise: *dress soberly when working at a bank*, *do not cause undeserved harm* etc, but rather defines a particular class of inference patterns. Using explicitly normative terminology such as "employees *should* wear neckties" serves to express endorsement of particular patterns of inference:

> Different patterns of inferences should be understood as corresponding to different sorts of norms or pro-attitudes. [7, p.90]

The most general or abstract way to define a norm is thus as a *subset* of the set of inferences available to an agent according to its propositional vocabulary and reasoning capabilities.

4 Data structures for reasoning about normative agents

This section outlines some data structures, at a fairly high level of abstraction, which could be employed in reasoning about the actions and commitments of a

normative agent, either by observing its behaviour and utterances or by directly querying it about the reasons for its actions. We begin by adopting some notation from d'Inverno and Luck's SMART framework [13] in the hope that this will facilitate comparison with more established approaches.

4.1 Basic definitions

I will follow d'Inverno and Luck up to the definition of an Agent, after which there will be some divergence.

The framework includes the following primitive types, where *Attribute* is the type of basic facts about the world:

$$[Attribute, Action]$$

Entities are not taken as primitives, but as bundles of attributes. What follows is a simple example of a Z schema [21], comprising a name (*Entity*), a section where variables are declared (the *signature*) and a *property* section.

Entity

$$attributes : \mathbb{P}\, Attribute$$

$$attributes \neq \varnothing$$

The Environment, *Env* is defined as some non-empty set of Attributes:

$$Env == \mathbb{P}_1\, Attribute$$

An Object an Entity capable of Actions. The notation here says that the *Object* schema includes the specifications of the *Entity* schema and extends it with additional statements.

Object

$$Entity$$
$$capabilities : Actions$$

$$capabilities \neq \varnothing$$

Objects do not have their own goals, so only act in furtherance of goals imposed from outside. If an Object is endowed with goals, it becomes or instantiates an Agent:

$$Goal == \mathbb{P}_1\, Attribute$$

Agent

$$Objects$$
$$goals : \mathbb{P}\, Goal$$

$$goals \neq \varnothing$$

4.2 Definitions for norm-conformant agents

In order to be able to talk about the reasoning capacities of normative agents, I define a few new types.
A Proposition is a bundle of attributes which may or may not be true in a given situation. It appears that a Proposition is denotationally equivalent to an Entity, but they will be differentiated by their different roles in agent schemas.

$$Proposition == \mathbb{P}_1\ Attribute$$

An Inference is an ordered triple involving an Environment, a set of Propositions constituting the premises of an argument, and a Proposition as the conclusion.

$$Inference == (Env \times \mathbb{P}\ Proposition \times Proposition)$$
$$Inferences == \mathbb{P}\ Inference$$

If we observe an Agent carrying out Inferences, we may call it a RationalAgent.

RationalAgent

$$Agent$$
$$inferences : Inferences$$
$$inferences \neq \varnothing$$

As argued in the previous section, norms are essentially constraints on inferences, thus they delimit the class of logically possible inferences an agent may carry out. A simple way to represent this is to define the *norms* which an agent adheres to as a subset of the *inferences* of which it is capable.

NormativeAgent

$$RationalAgent$$
$$norms : Inferences$$
$$norms \neq \varnothing$$
$$norms \subseteq inferences$$

This basic framework will of course need to be extended in various ways, in particular to model the *sharing* of norms within an agent society.

5 Conclusions

This extended abstract has considered normative agency in terms of the accountability of agents, where (following [8]) agents can be held accountable for their actions by being sanctioned for deviant behaviour, or by being required to give an account of the reasons and justifications for the action. The account has drawn on recent work in linguistic and social philosophy by Brandom and

Heath. I have proposed a clear distinction between norms and goals, characterising norms as constraints on reasoning which govern both the selection of appropriate goals and their justification after the event. The discussion has been conducted in rather general terms, though I have tried to indicate how it could be made more precise with the aid of the SMART framework. Future work will aim to extend this formalisation, in the hope that this will not only inform the design of normative software agents but can feed back into the philosophical arena by sharpening up the conceptual framework.

References

1. Bratman, M.: Intentions, Plans and Practical Reasoning. Harvard University Press (1987)
2. Austin, J.L.: How to do things with words. Oxford University Press, Oxford (1962)
3. Searle, J.: Speech Acts. Cambridge University Press, Cambridge, UK (1969)
4. Kibble, R.: Speech acts, commitment and multi-agent communication. Computational and Mathematical Organisation Theory **12** (2006) 127 – 145
5. FIPA: FIPA communicative act library specification. Technical Report SC00037J, Foundation for Intelligent Physical Agents, Geneva, Switzerland (2002) Specification dated 2002/12/03.
6. Brandom, R.: Making it Explicit. Harvard University Press, Cambridge, Massachusetts and London (1994)
7. Brandom, R.: Articulating Reasons: An Introduction to Inferentialism. Harvard University Press, Cambridge, Massachusetts and London (2000)
8. Heath, J.: Communicative Action and Rational Choice. MIT Press, Cambridge, Massachusetts and London (2001)
9. Kibble, R.: Reasoning about propositional commitments in dialogue. Research on Language and Communication (2006)
10. Habermas, J.: Theory of Communicative Action, vols 1 and 2. Polity Press, Cambridge, UK (1984)
11. Habermas, J.: On the Pragmatics of Communication. Polity, Cambridge, UK (1998) Edited by Maeve Cooke.
12. Singh, M.: A social semantics for agent communication languages. In: Proc. IJCAI'99 Workshop on Agent Communication Languages. (1999) 75–88
13. d'Inverno, M., Luck, M.: Understanding Agent Systems. Springer Verlag, Berlin Heidelberg New York (2004)
14. Lewis, D.: Convention. Blackwell, Oxford (1969)
15. Osborne, M., Rubinstein, A.: A Course in Game Theory. MIT Press (1994)
16. Lopez y Lopez, F., Luck, M., d'Inverno, M.: A normative framework for agent-based systems. In Boella, G., van der Torre, L., Verhagen, H., eds.: AISB'05: Symposium on Normative Multi-Agent Systems, University of Hertfordshire (2005)
17. Davidsson, P., Johansson, S.: On the potential of norm-governed behaviour in different categories of artificial societies. In Boella, G., van der Torre, L., Verhagen, H., eds.: AISB'05: Symposium on Normative Multi-Agent Systems, University of Hertfordshire (2005)
18. Walton, D., Krabbe, E.: Commitment in dialogue. State University of New York Press, Albany (1995)
19. Amgoud, L., de Saint-Cyr, F.D.: Towards ACL semantics based on commitments and penalties. In: Proceedings of ECAI 2006, Riva del Garda, Italy (2006)

20. Hamblin, C.: Fallacies. Methuen, London (1970)
21. Spivey, J.M.: The Z notation: a reference manual. 2nd edition. Prentice Hall (1994)