The Knowledge Engineering Review, Vol. 32, e18, 1–31.
doi:10.1017/S0269888917000169

Engineering the emergence of norms: a review

CHRIS HAYNES, MICHAEL LUCK, PETER MCBURNEY, SAMHAR MAHMOUD, TOMÁŠ VÍTEK and SIMON MILES

Department of Informatics, King's College London, Strand, London, WC2R 2LS
e-mail: christopher.haynes@kc.ac.uk, michael.luck@kcl.ac.uk, peter.mcburney@kcl.ac.uk, tomas.vitek@kcl.ac.uk, simon.miles@kcl.ac.uk

Abstract

Complex systems often exhibit emergent behaviour, unexpected macro-level behaviour caused by the interaction of micro-level components. In multiagent systems, these micro-level components may be autonomous agents and the emergent behaviour may be expressed as norms—patterns of behaviour that arise among the agents in response to their environment and each other. These emergent norms may be beneficial (e.g. by encouraging cooperative behaviour), or detrimental, but in either case it is useful to recognize these norms as they emerge and either encourage or discourage their establishment. We term this process engineering the emergence of norms and have identified three steps: the identification of a possible norm, evaluation of its benefit and its encouragement (or discouragement). This paper is an attempt to provide a survey of existing research related to these steps. We also provide an analysis of the approaches based upon their suitability for a variety of normative systems: we examine the requirements for agents to have autonomy over their choice of norms, the degree of observability required in the system, and the norm enforcement methods. The paper concludes with an discussion of open issues.

1 Introduction

Recent times have seen the advent of large-scale networks of distributed devices that gather data and communicate among themselves in order to solve problems in application domains as diverse as logistics, healthcare, tourism and manufacturing. This has led to the development of closely related paradigms such as the Internet of Things (IoT), ambient intelligence and ubiquitous computing. Such systems bring about a need for flexible, decentralized control mechanisms that can cope with autonomous software agents, with heterogeneous capabilities and goals, working within dynamic environments.

Control mechanisms are required for two reasons: first, to discourage behaviour that can harm the system or other agents; and second, to allow the agents to coordinate their behaviour efficiently. The members of such networks may be autonomous with respect to the network itself, with heterogeneous capabilities and goals, and controlled by different entities. Rigid control, or regimentation, may not be possible—resources may not be available to strictly monitor every agent, and it may not be possible to always prevent harmful actions. Even if it is possible to regiment behaviour, it may not be desirable to limit the agents' autonomy and creativity, since this may reduce the usefulness of the system, especially where agents are required to overcome problems not foreseen by the system designer, perhaps using capabilities that were not envisaged when the system was originally built.

While this issue is inherently complex, we do have examples of large-scale multiagent systems made up of autonomous, heterogeneous individuals from which to take inspiration—human societies. Behaviour within human societies is guided and controlled by norms, either informal social norms, such as table manners, or more formal legal norms, such as laws against murder or tax evasion. Norms guide the societies' members by indicating what behaviour is considered acceptable, and what is not, as well as

providing a basis for expectations about the behaviour of others that can prove to be useful when coordinating behaviour. While human beings are not rigidly controlled by these norms, those caught violating them usually suffer some kind of punishment, such as a loss of reputation when violating etiquette, or imprisonment for committing murder. These punishments, as well as the possible benefits accrued from improved coordination, encourage humans to adhere to the norms of the society.

Legal norms are explicitly designed to control behaviour, for example, laws are written and imposed by governments, but social norms emerge within groups of people from interactions either between agents, or between agents and their environment, in an unplanned, bottom-up fashion. They are not designed like laws, and the punishment for violating a social norm is usually more subtle and uncertain than the punishment for breaking a law. Some social norms exert as much, if not more, control over human behaviour as legal norms—for example, when duelling to defend one's honour was considered an obligation amongst army officers, many were prepared to violate the legal norm prohibiting it in order to comply with the social norm. Although the direction of this control is not consciously and explicitly designed, many social norms do arise in response to social problems of coordination and cooperation, and some theories of social norm creation view them as equilibria of behaviour.

Returning from the example of human societies to the problem of controlling the behaviour of agents within a very large-scale open network such as the IoT, the emergent nature of social norms provides an opportunity (as well as challenges). Given that social norms may solve societal problems and that explicitly designing norms can be extremely hard (Shoham & Tennenholtz, 1995), it may be advantageous to use emergent social norms to control behaviour rather than trying to design and impose norms. Of course, where all of society already complies with a certain social norm there is little, if anything, to be done. However, if only part of a society adheres to a useful norm, then efforts can be made to spread that norm to the rest of society in order to increase the benefit. For example, drivers who obey speeding laws may do so as much through social pressure (a norm) as through fear of getting caught, whereas persistent speeders may feel no such pressure, and hence no guilt or shame at speeding (De Pelsmacker & Janssens, 2007). Spreading the anti-speeding norm to all drivers may therefore decrease the incidence of speeding and make the roads safer.

This poses a number of challenges. First, social norms are often not explicit, but may only manifest as patterns of behaviour, and so it may be hard to detect what is or is not a social norm. Second, there is no guarantee that a social norm will be beneficial; history abounds with social norms harmful to individuals and societies. Third, encouraging social norms to spread is not straightforward.

In this paper, we review the literature concerning making use of emergent norms in complex agent-based systems to serve the goals of those systems, or in other words, the *engineering* of emergent norms. This engineering has three main steps: detection of possible emergent norms, evaluation of those norms in terms of the benefit they bring to individuals and the system as a whole, and encouraging the spread of beneficial norms throughout the system (or discouraging detrimental norms). With this in mind, we examine the state of the art in each of these steps and seek to identify possible future research from the perspective of this engineered approach.

We do not solely concentrate on norms, but also consider conventions and emergent behaviour in general especially in the detection and evaluation steps. Norms are generally considered to include a deontic aspect, but this aspect need not arise until agents begin to encourage (or discourage) the behaviour. At least in the detection and evaluation steps it would be more precise to say that we consider proto-norms (behaviour that may become a norm).

This paper is motivated by the recent increased interest in large-scale complex systems, such as those encompassed by the IoT paradigm (Atzori *et al.*, 2010). Within this paradigm, it is proposed that large-scale networks of distributed 'things' will communicate together and share data in order to solve problems in application domains as diverse as logistics, healthcare and tourism. Such systems will inevitably exhibit emergent behaviour—macro-level patterns of behaviour that are caused by the interactions of the micro-level individuals (or agents) within the system. While this emergent behaviour may be detrimental to the system, it may also be beneficial, especially with sophisticated agents, as they may find solutions that were not predicted by the system designers—indeed, the problems to which these solutions apply may themselves be unexpected ones not predicted by the designers. If such systems are to work efficiently, detrimental behaviour must be discouraged and beneficial behaviour encouraged.

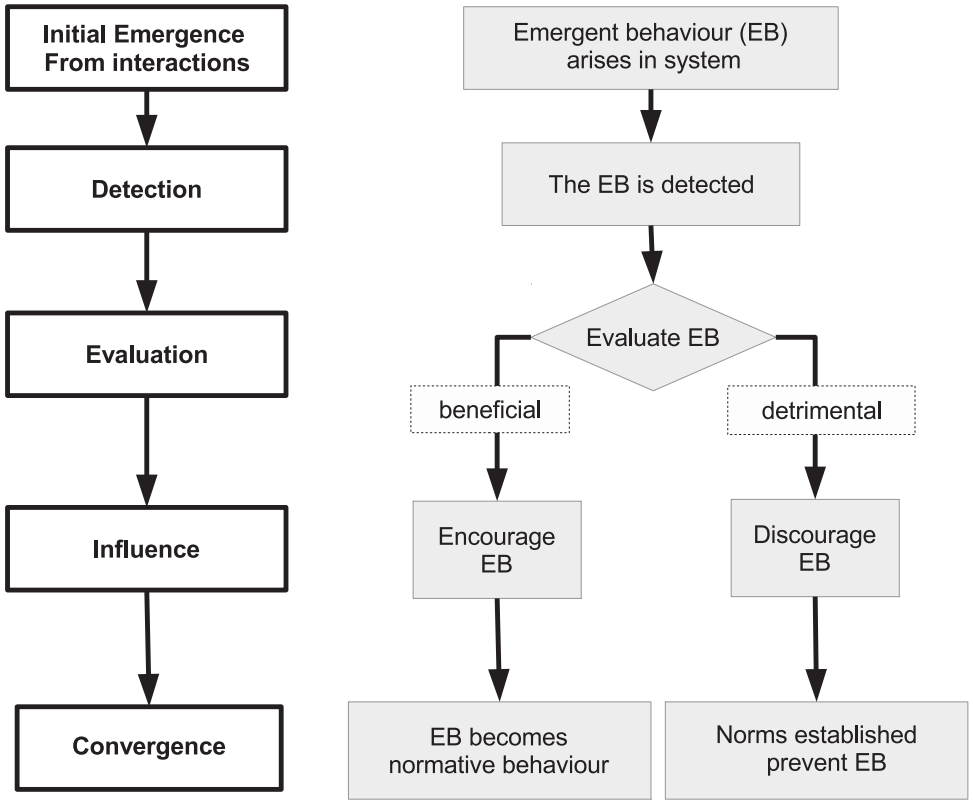

Figure 1 The main steps of harnessing emergent behaviour

With respect to the detection step, in order to maintain the focus of the paper, we restrict the scope of the paper to reasoning mechanisms that identify norms and emergent behaviour. In particular, we do not review the important related issues involved in gathering data in practical, distributed, multiagent systems, such as privacy, security and authority. We do, however, analyze approaches based upon their requirements for observability within the system (see Section 5). We also group them into agent-based and system-based approaches, since this can determine the nature of how data can be gathered and stored.

Figure 1 shows a conceptual view of the steps required to harness and control emergent behaviour in a system. Assuming that emergent behaviour has arisen in the system, the first step is to detect this behaviour. Once detected, some evaluation is required so that the behaviour can be judged either beneficial or detrimental. In the next step, the agent behaviour must be influenced to either encourage beneficial behaviour, or discourage detrimental behaviour. This leads to the establishment of a norm.

We present the process here in a linear fashion, concerning only a single norm from initial emergent behaviour to eventual convergence, because, in this paper, we are only concerned with these steps. Once a norm is established, it may change as agents encounter other circumstances due to the environment changing or the modification of their goals and capabilities. In particular, after convergence a norm must be periodically re-evaluated to ensure that it is still valuable, and monitoring must detect new emergent behaviour that arises. In addition, even norms that have converged may decay if agents choose not to comply, so the influence step may need to be re-visited if the norm is still considered useful. In short, we have simplified the complex cycle and sub-cycles of the norm life-cycle to maintain clarity.

To be specific, we reviewed the literature for research which seeks to answer the following questions:

- How can an emergent norm or pattern of behaviour be detected in a multiagent system?
- How can the emergent norm be evaluated with respect to the needs of individual agents and the multiagent system as a whole?
- How can valuable norms be encouraged, and harmful ones be discouraged?

We intend this paper to be a useful resource for designers of normative multiagent systems interested in detecting and harnessing emergent norms in those systems. To this purpose, we describe the main approaches taken for answering these three questions within the literature and analyze these approaches with respect to the system properties to which they are applicable.

There are two broad strands of research on norms in multiagent systems: works studying the design of norms for some purpose, and works studying the emergence and spread of norms. The former strand largely focusses on norms that are explicitly designed by the system owner to fulfil a certain known purpose, in the same way that human laws (legal norms) are created by some law-giving entity to satisfy some perceived need. Work in this strand includes mechanisms to monitor norm compliance, and the design and evaluation of norms. The latter strand encompasses work examining how norms emerge and spread in different types of network, and how agents recognize and choose to adopt the newly emergent norms. We draw upon works from both strands in this survey, as well as some relevant works outside the norms literature.

This paper is structured as follows: first, we discuss existing survey papers in the field of norm emergence and relate them to this work in order to highlight our contributions (Section 2); second, we review the definitions of conventions, norms and laws, and examine the problems of using social norms as a form of control (Sections 3 and 4); third, we examine current work both on norm identification, and on identifying emergent properties in general (Section 6); fourth, we review the existing work on norm evaluation, from both an agent and system perspective (Section 7); fifth, we review the research into encouraging the convergence of norms in multiagent systems (Section 8); and finally, we identify possible future research directions in Section 9.

2 Existing surveys

There are several other survey papers that examine research in norm emergence in the context of multiagent systems. In this section, we briefly describe these works and note the similarities and differences with this paper: first, to highlight our unique contributions beyond merely including work performed after the publication of these survey papers; and, second, to detail where these works cover valuable work that is outside the scope of this survey.

Criado *et al*. (2011) present a review of normative multiagent system research as part of a discussion of open issues in the field. In particular, they provide a very useful survey of norm representation and implementation approaches that we do not cover in this paper. They briefly cover the topic of norm emergence and the factors that affect it, but do not examine it in any depth.

Hollander and Wu (2011) provide a high-level overview of research into normative agent-based systems. They briefly discuss emergence and the effect of punishment and norm spreading upon this emergence, however, as in Criado *et al*. (2011) the breadth of their work precludes an in-depth review or analysis of the topic. In contrast, we seek to cover these issues in much greater depth, as well as including factors such as norm evaluation and identification that are related to our concept of engineering norms for specific system purposes.

A survey more focussed upon norm emergence is provided by Savarimuthu and Cranefield (2011). They propose a norm life-cycle with creation, identification, spreading and enforcement leading to the widespread adoption of a norm. Instead we use the perspective of deliberate engineering of norms, and so explicitly consider the evaluation of the benefit a norm provides and consider enforcement as only one of the factors that influence the spread and adoption of a norm throughout a society. In addition, with respect to identification, their focus is purely upon norms, whereas we examine the identification of emergent behaviour patterns that are not yet norms—so that this behaviour can be engineered into a system norm. Also, we examine work that is outside of the multiagent field, but is relevant to emergent behaviour, in particular in Section 6.2.

While these other surveys detail some of the factors that influence the emergence and spreading of norms in a multiagent system, for example, norms about norms (metanorms) and network topology, we strive to show how those factors may be used in the engineering of the system to promote or hinder the emerging norms. This focussed, coherent approach from a norm engineering perspective is the main contribution of this paper.

3 Conventions, norms and laws

In this section, we discuss the differences between conventions, social norms and laws. There are no specific accepted definitions of these terms, and authors may use them differently from here[1], however, we attempt to give the most common definitions. We note that the boundary between convention and social norms, especially in human societies, is somewhat fuzzy.

A convention is a stable pattern of behaviour, or equilibrium, within a society. There is no deontic aspect to a convention—it merely describes what *is*, rather than *ought* to be (Conte & Castelfranchi, 1999). A convention does not impose an obligation, and violating a convention incurs no ill effects beyond those naturally derived from acting contrary to it. In other words, there is no punishment from other actors within the society if one chooses to disregard a convention. Gibbs (1965) proposes a typology of all forms of norms from a sociological perspective. He uses three attributes: collective expectations, collective evaluations and reactions to behaviour. He uses the low probability of violations leading to sanctions as a means of distinguishing conventions from other norms.

A norm, in contrast to a convention, imposes an obligation to act, or not act, in a particular way, and the violation of a norm incurs the risk of punishment if the violation is detected. For a *social* norm, the punishment is enacted by other members of the society who choose to uphold the norm. Conventions may evolve into social norms, especially if the society values conformity and is prepared to sanction those who do not conform—in this way, a pattern of behaviour that begins as a mere convention may come to be regarded as a social norm.

According to Bicchieri, social norms are the 'grammar of society' that 'specify what is acceptable and what is not in a social group' (2006). She notes that norms create expectations as well as obligations, and in her definition a social norm only exists if a large majority of a group expect that most members of that group will comply with the norm and act in the way it obliges.

Axelrod uses a behavioural definition in his seminal paper: 'A norm exists in a given social setting to the extent that the individuals usually act in a certain way and are often punished when seen not to be acting in this way' (1986: 1097). It is notable that he includes the notion of a *measurable* degree of norm existence in a society, rather than defining a norm as either existing or not existing in a society. This allowed him to examine the growth and decay of norms. It is more common to see this degree of existence referred to as the degree of *convergence* within a society (Walker & Wooldridge, 1995).

In human societies it is common to formalize many of the social norms into a set of legal norms (or laws). For example, social norms against violence are commonly formalized into laws against assault, manslaughter, murder, etc., with specified contexts and punishments. This formalization has a number of advantages over simply relying on social norms. Hart (2012) identified three defects of a body of rules made up of social norms as seen in a pre-legal society: uncertainty, stasis and inefficiency.

First, social norms are not always easy to identify (as we shall see in Section 6), and so members of a society may not agree on the extent and meaning of the requirements of a norm. This can lead to disagreements about when a norm is applicable and what punishment is suitable for a specific violation. This uncertainty can impact both cooperation and future planning, where this depends upon another's interpretation of a norm.

Second, social norms, as opposed to laws, are usually spread by diffuse social pressure and consensus. As a consequence, they are often very slow to change, and importantly cannot usually be changed deliberately. Although individuals can seek to persuade others in order to change their norms, this is a slow process unless the persuasion is backed by force or offers of reward. Even if laws are enacted to try to change a social norm, it is not usually a swift and predictable process. For example, Axelrod (1986) gives the example of duelling in which the power of the law was insufficient to change the social norm that obligated men to duel in defence of their honour[2]. In addition, while such norm changes are occurring, the society suffers from even more uncertainty over the status of norms than usual since individuals may not know whether others expect them to obey the old norm or the new one.

[1] Fortunately, the lack of accepted definitions leads most authors to make their semantics clear in each paper.

[2] Of course, there are examples of laws that do change behaviour rapidly, such as the seat-belt laws we describe in Section 8.

Finally, punishments are reliant upon social pressure and haphazard enforcement. With no official agency to enforce norms, they are an inefficient form of social control. Even the adoption of social norms that oblige witnesses to punish norm violators does not compensate for the lack of an official agency with the power to enforce the rules[3].

Hart (2012) proposes three meta-rules, the presence of which distinguishes a pre-legal society from one with a legal system. The *rule of recognition* specifies which norms are recognized as laws by the society. This addresses the uncertainty problem since it turns unwritten rules into codified laws. The *rule of change* specifies the way in which legal norms may be changed. This addresses the stasis problem as it provides a way to cope with dynamic environments by changing the laws of society. It reduces the need for a broad social consensus to a need for a consensus among those empowered to change the laws, and it allows rule changes to be coordinated across a society. Finally, the *rule of adjudication* specifies both who has the power to judge if a rule has been violated and how the process of judging and punishment is allowed to be carried out. This addresses the inefficiency issues, and also helps reduce uncertainty over-punishment. Note that it is not uncommon for a legal system to suffer from uncertainty, stasis and inefficiency, if their meta-rules are poorly designed or violated.

Within the computer science field, work on formalizing norms has proposed representing them as contracts (also known as *electronic contracts*) between agents to establish specified expectations of behaviour, often connected to explicit punishments if those expectations are not met. For example, Singh (2013) presents a model that uses norms, specified as contracts, to govern agent behaviour. Their model includes, among other things, notions of authority, punishment and the power to change the norms. Such normative electronic contracts require monitoring for violation, and Modgil *et al.* (2015) present a monitoring framework based on the observation of agent behaviour. Work on electronic contracting mostly focusses upon norm devised in a top-down fashion, rather than the emergent norms that concern us in the paper. However, given the growing interest in large-scale, decentralized, self-adapting systems, it seems likely that there will be a need for such systems, or autonomous software agents within those systems, to generate their own contracts based upon useful norms that emerge during runtime. We will examine the identification, evaluation and encouragement of such emergent norms in this paper.

4 Emergent behaviour and emergent norms

We have discussed the range of definitions of the term social norm in the previous section. In this section, we examine emergent behaviour in general, what it means for a norm to emerge in a society, and discuss the difference between an emergent and an established norm.

At the simplest level, to *emerge* can mean to appear, or arise, from a process, especially in an unexpected way. However, in the literature, emergent behaviour has come to be characterized as more than simply behaviour that appears unexpectedly in a system. There are many definitions of emergence and emergent behaviour which we do not cover here. (Deguet *et al.* (2006) give a brief survey.) Instead, the following broad, informal definitions suffice to show what we mean by emergence.

Bedau (1997) defines emergent phenomena[4] as being generated or derived by underlying processes, but autonomous from those underlying processes. Implicit in the definition is the presence of multiple layers in a system: the micro-level, constituting the individual components, and the macro-level constituting an abstraction of the system as a whole. For example, a nation state is made up of individuals which have properties (such as age, and incomes), but the state itself also has macro-level properties, such as gross domestic product (GDP) and inflation rate. Wolf and Holvoet (2005) define emergence as coherent behaviour at a macro-level that arises dynamically from micro-level interactions while being novel with respect to the micro-level components of the system. They further suggest that emergent behaviour is decentralized and robust with respect to the replacement of individuals—in other words, the behaviour is

[3] Norms concerned with how agents should react to norm violations are referred to as metanorms within the literature (Axelrod, 1986). We examine research on using metanorms to influence norm emergence in Section 8.1.

[4] Bedau terms the emergent phenomena we are concerned with as weak emergence to distinguish it from a metaphysically inconsistent version he terms strong emergence.

stable even if some individuals are removed or replaced. Implicit in these definitions is a lack of intentionality and deliberation on the part of the micro-level components to bring about the macro-level effects[5].

When discussing social norms the use of the term emergence can cause confusion, since it is common in the literature to use the term to describe the process by which a norm becomes established throughout the society. Further, a norm is said to have *emerged*, or to have become *established* when a sufficiently high percentage of the population complies with the norm. Whereas Villatoro *et al.* (2011b) define a robust convention as requiring 100% convergence, most definitions of emergence do not expect such complete uniformity, but instead require 90% convergence (Kittock, 1994; Shoham & Tennenholtz, 1997; Delgado, 2002).

Since social norms typically arise from the interactions of individuals with each other and their environment in an unplanned way, they are clearly emergent properties of a society. Even if a norm originates from a single agent deliberately trying to change the behaviour of others via the establishment of a norm[6], the spread of that norm via diffuse social pressure can be seen as emergent. However, there is no assumption that emergent behaviour is global (or even predominant) throughout a system. In this paper, we use *emergent norm* to describe a norm that has arisen as an emergent property, regardless of whether it is followed by a majority of the population, or indeed whether it is recognized as a norm at all. An *established norm* is one that has reached a stable state where the majority of agents comply with it. This is not to say that established norms cannot collapse, especially if circumstances change.

5 Analysis dimensions

As well as describing a representative selection of the approaches useful in the steps of engineering emergent norms, we seek to analyze those approaches with respect to their suitability for different types of multiagent system. With this in mind, in this section we discuss certain properties of normative multiagent systems that influence that suitability. Specifically, we examine the following properties:

- autonomy;
- observability;
- enforcement.

While *autonomy* is usually considered a necessary property of an agent, there are different aspects and degrees of autonomy (Castelfranchi & Falcone, 2003). In particular, we consider whether different approaches require that agents in a system have *norm adoption autonomy* (the power to choose their own norms). Note that a lack of norm adoption autonomy does not mean an agent must obey the norms, merely that it cannot choose them. For example, a group of agents in a coalition of peers may decide upon their own norms, so they have norm adoption autonomy. In contrast, a set of agents designed and built by a single organization to fulfil a task may have norms imposed upon them, and while they may choose to violate the norms under certain circumstances they cannot choose to generate and adopt new ones.

Some approaches depend upon agents being able to *observe* their neighbours' actions, and the results of those actions, either directly or mediated through a reputation system. In particular, agent-based approaches rely on agents being able to observe their peers. In some systems, this is impossible due to privacy or security concerns, or environmental factors. Other approaches require that some centralized mechanism is able to gather and store global information about agent activity. This may be impossible for various reasons, for example: privacy issues may not allow a central system to gather this information about individual agents. Also, the environment or the size of the system may preclude gathering this data in real-time (e.g. if logs are gathered and examined only periodically). In some domains, systems may use a multiagent approach specifically because a global approach is not possible (Sycara, 1998). Therefore, the choice of approach can be restricted by observability.

Norms may be *enforced* by peers or the system itself. The choice of enforcement method, and its effectiveness, is determined by the nature of the system and the agents. For example, in a peer-to-peer file

[5] Mintzberg and Waters (1985) make this explicit in their definition of emergent organizational strategies.
[6] Such an agent is known as a norm entrepreneur (Finnemore & Sikkink, 1998).

Table 1 Example multiagent systems and their properties

Systems	Norm autonomy	Observability	Enforcement	References
Ambient home network	No	System	System	Campillo-Sanchez and Gomez-Sanz (2015)
Sociotechnical system	No	Peers	System	Singh (2013)
Wireless mobile grid	Yes	Peers	Peers	Balke *et al*. (2012)
Open source software projects	Yes	System, peers	Peers	Savarimuthu and Dam (2013)

sharing system, agents may be able to punish peers by refusing to share files with them, but if agents are able to access the system anonymously then this may be impossible. In a private marketplace system, peers may be unable to punish each other, but the system controller can bar access to norm violators. Note that an agent must be able to observe its peers if it is responsible for norm enforcement. Peer enforcement may be restricted by the nature of the system, but also by system policy (e.g. peers may not be allowed to enforce norms on an ad hoc basis).

Table 1 shows four examples of multiagent systems and their properties to illustrate the dimensions of our analysis. Note that we consider specific systems, or proposed systems, here—not all instances of a wireless mobile grid will have the properties detailed here.

An ambient home network consists of networked devices, both mobile and static, within a home that share data to perform tasks. Campillo-Sanchez and Gomez-Sanz (2015) propose such a network, using norms to constrain agent behaviour, to facilitate independent living for a user suffering from Parkinson's disease. An example norm would oblige a TV remote control app to make its buttons bigger if the user is having a tremors episode.

Singh (2013) describes a sociotechnical system, the Ocean Observatories Initiative (OOI), a multi-stakeholder system made up of autonomous computational and physical resources designed to coordinate oceanographic research and monitoring. Singh proposes the use of norms to constrain the behaviour of the agents. There is no global view, but agents interact with and may observe their peers. Norms are enforced by the central OOI organization. An example norm is the prohibition of publishing shared data without permission from the owner of the data.

Wireless mobile grids are a proposed mechanism to create *ad hoc*, decentralized networks for sharing resources across mobile devices (Fitzek & Katz, 2007). Balke *et al*. (2012). proposes using norms to ensure users do not indulge in selfish behaviour and they note that emergent norms may develop due to interactions between agents (Singh *et al*., 2013). In such a system, agents may observe the behaviour of their peers and are responsible for enforcing norms. An example norm could prohibit an agent from accessing resources without offering resources to other agents in the network.

Savarimuthu and Dam describe the issue of norms emerging within open source software projects, and the notion of extracting and studying those norms by means of the repositories that hold developer discussions and information about project updates and bug reports (Savarimuthu & Dam, 2013; Singh *et al*., 2013). In this domain, the agents are human users and developers. Since the repositories are public, there is system-wide observability. Enforcement of the norms may be done either by the system or by the agents. An example norm may be the obligation to make sure that checked-in code compiles: depending on the repository this may be enforced by the system automatically sending a warning message, or by peers making their displeasure known via manual messages to the norm violator.

6 Identifying emergent norms

In this section, we examine how emergent norms may be identified, both by agents within a society, and by system level approaches. Since emergent patterns of behaviour can become norms, and since we are interested in influencing behaviour that has not necessarily become established as a norm, we do not focus

solely on norms, but also consider research on identifying emergent behaviour in general. This is particularly true for Section 6.2, where we discuss some approaches that assume no deontic aspects to the behaviour, but merely look for the patterns (such as rule-mining).

6.1 Agent level identification

The approaches in this section are all based upon individual agents reasoning about the existence of norms. Although they may observe other agents and communicate with them, the actual inference is performed by a single agent. Existing research in norm identification[7] mostly concentrates on newcomer agents ignorant of the norms within a society learning those norms. These are, of course, not new norms emerging in the system. However, we suggest that these approaches should also be able to detect newly emerging norms before they are fully established since the case where an agent is trying to infer an existing norm (that it is unaware of is) is similar to the case where it is trying to infer an emerging norm. Therefore, it is those approaches that we focus on in this section.

Savarimuthu and Cranefield (2011) identify three ways that an agent can identify norms. First, experiential methods, where the agent performs behaviour and is punished if it violates a norm; second, communication methods, where the agent is explicitly informed of a norm by other agents, or system artefact (such as a list of rules); and third, where it infers the existence of a norm by observing other agents either violating a norm and being punished, or consistently avoiding some possible behaviour.

For example, consider a norm against walking on the grass in a park. If an agent unfamiliar with the norm enters the park, it may learn of the norm in the following ways. If it steps on the grass and is punished, then it has learned experientially. If it reads a sign saying, 'Do not walk on the grass', it has learned via communication. If it sees another agent being punished after walking on the grass, it has learned by observation. Finally, if it notices that no other agent walks on the grass then it may infer that it is prohibited by a norm, also by observation.

Andrighetto *et al*. (2010a) simulate a norm recognition mechanism based on a cognitive approach. Agents receive information about possible norms in the form of either messages from other agents or by observing their actions. These messages and observations are analyzed for either deontic ('You must do X') or normative value components ('Not doing X is bad'): if these are present then the agent stores the behaviour as a possible norm. As more such observations are recorded, the salience of the possible norm is increased—that is, it is deemed more likely to be an important norm. If multiple conflicting norms apply to a situation then the most salient is complied with. If agents have personal goals and desires, these can also be weighed against the salience of the norms. While they evaluate their approach via agent-based simulation they do not propose an actual mechanism for analyzing the messages and observation for deontic or normative value content. Instead, they assume that the agent is able to perform this analysis using some undefined method.

One observational approach is to detect the use of signalling behaviour, where some agent takes an action to show a norm has been violated (e.g. punishing another agent), to allow the inference of the existence of a norm. Savarimuthu *et al*. (2010, 2013a) propose norm identification mechanisms for identifying obligation and prohibition norms in agent societies. When an agent perceives a signalling event it invokes a norm inference component to determine whether the event may have occurred as a result of the violation of an unknown norm. The inference itself is a three step process: first, the agent records sequences of events that it observes; second, when the agent perceives a signalling event it extracts the sequence of events related to the sanctioned agent that preceded the signal; finally, the agent generates a set of candidate norms using a rule-mining algorithm. These candidate norms are then verified by communicating with other agents (who are presumed to know the norms of the society).

There are three obstacles to using violation signals to infer norms: first, norm violations may be rare, second, the signals may occur sometime after the violation, and finally, a newcomer may not recognize the signalling behaviour at all. For example, in a law-abiding society where only a very few agents violate the norms, newcomers may rarely witness violation signals.

[7] Within the literature it is also referred to as norm recognition or norm learning.

Instead of inferring norms from the observation of violations, an agent could learn by observing compliant behaviour and comparing that with possible ways of reaching the same goals. By detecting actions not taken towards goals, it may be possible to infer prohibitions or obligations. For example, if there are five possible roads by which an agent could travel from point A to point B, and one road is consistently avoided, then there may be a norm prohibiting that road. Oren and Meneguzzi (2013) propose such a method using two components: a plan recognizer and a planner. The agent first observes the behaviour of an agent in the society and uses the plan recognizer to infer the overarching goal of that agent. Once the goal has been identified, it uses the planner to generate a set of possible plans by which that goal can be achieved. Sets of all possible obligated and prohibited states are generated and over a series of observations states are removed from each set, and states are added to a set of potential prohibited states. States that are entered by agents are deemed to be not prohibited and removed from the potential prohibited set; states that are not entered are removed from the set of potential obligations.

While their initial approach assumes that all agents comply with the norms, Oren and Meneguzzi extend it to allow the possibility of violation. Instead of monotonically removing states from the potential obligation and prohibition sets they allow for a certain ratio of violation to compliance. In simple terms, they keep a count of possible obligations and prohibitions and also a count of violations of those possible norms. If the ratio of compliance to violation is above a specified threshold then the norm is inferred.

Cranefield *et al*. (2015) combine both the observation of signalling behaviour and plan recognition in order to infer norms. They use Bayesian reasoning to generate a list of norms ordered by likelihood. Another combination approach is proposed by Mahmoud *et al*. (2012a) who present a norm detection framework that uses data mining techniques on three sources of data: system logs of agent activity, communication with local agents, and observations of agent activity. The security policies of the system determine which of the three sources an agent can actually use.

6.2 System level identification

Existing work on the detection of emergent properties in complex systems is not specifically focussed on norms, but in so far as social norms are derived from emergent patterns of behaviour, the techniques are appropriate. Emergent behaviour in complex software systems, such as computer networks, or load-balancer applications, is a practical problem that has been investigated. Mogul (2006) categorizes the most likely behaviour and causes in such systems. In such systems, the presence of emergent behaviour is detected largely by examining the difference between the actual and expected behaviour. Parunak and VanderBok (1997) discuss techniques to detect possible emergent behaviour in distributed control systems, including Fourier analysis to distinguish periodic events from random noise. They propose using agent-based simulation to examine possible causes of emergent behaviour within a system.

Moving into more general detection of emergent behaviour, we examine three different strands of research: variable-based detection, event-based detection and data mining. Variable-based detection relies on macro-level variables representing system properties. Changes in these macro-level variables may constitute emergent behaviour. For example, the GDP of a nation is a macro-level variable. GDP depends upon the behaviour of micro-level components of the nation (humans), and changes to the GDP are not directly calculable from the interaction of those components. Seth (2008) proposes a statistical approach to measure the degree of emergence of a macro-level variable in relation to a set of micro-level variables. He calculates the degree of emergence based upon the extent to which the macro-level behaviour is both caused by and autonomous from the micro-level behaviour. Kubík (2003) presents a grammar-based approach where macro-level and micro-level properties are expressed as two different grammars. Emergent behaviour is defined as those macro-level properties that cannot be derived from summation of the micro-level properties. There are two problems with variable-based detection: it relies on the aggregated behaviour of the individual micro-level components (albeit in an unpredictable way), and there is a need to specify the macro-level variables of interest and to choose which of them to monitor, which often requires expert domain knowledge. The aggregation means that if emergent properties are detected, it can be impossible to work out which micro-level actions or interactions caused them. Norms are an intrinsically

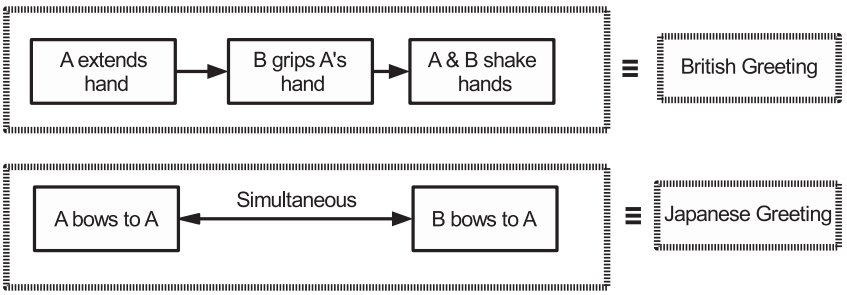

Figure 2 Two example norms represented as complex events (after Chen *et al*., 2009)

micro-level phenomenon, although their effects may not be, so this represents a fundamental problem with this approach.

Event-based detection solves this problem by allowing macro-level behaviour to be defined in terms of simple micro-level events. Chen *et al*. (2009, 2007) propose an event-based formalism of emergent behaviour for specifying and detecting such behaviour in an agent-based system. They define a complex event as a set of temporally, or spatially, related simple events, where a simple event is a state change in the system (typically caused by an agent performing an action). A complex event can also be made up of related complex events. Temporal relationships specify when one event occurs in relation to another and may include simultaneity, at some time before, at some time after, and immediately after; spatial relationships specify where one event occurs in relation to another. This formalism also allows variables to be specified as part of a relation. Sets of complex events are specified to represent different types of possible emergent behaviour. These can then be watched for and detected as the agents perform actions that change the system state. As complex events are ultimately defined in terms of simple events, the cause of the emergent behaviour can be determined (unlike with variable-based detection in which behaviour is aggregated).

Figure 2 shows two norms expressed as complex events: a British greeting (handshake) and a Japanese greeting (a mutual bow). A British greeting is defined as one agent extending a hand, a second agent gripping it, and a mutual handshake. A Japanese greeting is defined as two agents simultaneously bowing to each other. In this way, the two greeting norms can be specified and observers can look for instances of either phenomena.

One disadvantage of this approach is that possible emergent behaviour must be specified in order to monitor the system for the behaviour. Therefore, some entity must derive a set of such behaviours *a priori*. One possibility is to specify high-level behaviour that, if identified, can be used as a starting point for more detailed investigations.

Data mining can be used to identify patterns within large quantities of data. Of particular relevance to detecting emergent behaviour of agents is association rule-mining (Kotsiantis & Kanellopoulos, 2006) which extracts correlations, patterns and associations between items of data (such as logs of agent actions). This form of data mining has been used to formalize business processes and workflows (Van der Aalst *et al*., 2003) and so to extract structure from data. These processes represent the actions of entities within the business (e.g. human workers acting in response to customer requirements), and so it seems likely that these techniques could be used to detect emergent patterns of behaviour in multiagent systems.

6.3 Analysis

Table 2 summarizes the examined approaches and their requirements. The important system property with respect to emergent norm identification is, of course, observability, since one must be able to observe a behaviour in order to determine whether it constitutes an emergent norm. However, it would be possible for the agent level approaches to be implemented as part of a mechanism where the system uses specific monitor agents empowered to observe the other agents in order to identify emergent behaviour. Likewise, individual agents could make use of the event-based detection mechanisms proposed by Chen *et al*. (2009), even if system-wide observability was not possible.

Table 2 Summary of norm identification approaches

Approach	Observability requirements
Andrighetto *et al.* (2010a)	Peers
Savarimuthu *et al.* (2010, 2013a)	Peers
Oren and Meneguzzi (2013)	Peers
Cranefield *et al.* (2015)	Peers
Mahmoud *et al.* (2012a)	System/peers
Seth (2008)	System
Kubík (2003)	System
Chen *et al.* (2007, 2009)	System
Parunak and VanderBok (1997)	System

7 Norm evaluation

In this section, we examine the current approaches to evaluating a norm with respect to the benefit it provides. The notion that a norm can be evaluated presupposes that there is some metric against which it can be evaluated. If a norm has been designed towards some goal, then its value can be assessed by how it helps agents (or the system as a whole) meet that goal. However, when a norm emerges, it may not be straightforward to evaluate it. Importantly, this evaluation must take place from a specific perspective, since behaviour that is beneficial to one entity may not be beneficial to another. Possible perspectives include:

- an individual agent;
- the system designer or controller;
- a regulator, either within the system or external to it.

An individual agent may evaluate a norm in order to decide whether to adopt and comply with a norm, while evaluating from a system perspective can be used to decide whether the norm is appropriate for the system as a whole. In particular, because different entities can have different goals, each agent may evaluate a norm differently. In addition, individual agents may be unable to evaluate the long-term or indirect effects of a norm, especially if they only consider the impact of the norm upon their own behaviour.

In a system with overall goals, things are simpler. For example, in an open system designed to enable peer-to-peer file sharing, emergent norms can be evaluated based upon whether they facilitate efficient sharing and discourage free-riders. However, in open systems or societies where there is no overall consensus on goals, or where there are disagreements among the agents regarding the priorities of goals, different groups of agents may evaluate norms differently, and it may not be meaningful to speak of an overall evaluation. In such a case, a society that wishes to promote beneficial norms must use some kind of group decision-making mechanism, such as voting for which behaviour to encourage. A regulator of the system, may also have its own independent perspective. For example, the Bank of England regulates financial transactions in the United Kingdom that it is not itself a party to, and so its perspective on the financial norms and regulations is different to those of the banks (and other agents) within the system.

For both individual and system approaches, evaluation can encompass both the benefit and costs of adopting or complying with the norm, and those of not adopting the norm or of violating it if it already exists in the system. Benefits may accrue from increased cooperation or coordination, both on a personal and system level. Costs of not adopting or complying with an existing norm may include punishments, but also indirect costs due to reduced coordination (e.g. interfering with others that are following the norm). Adopting a norm also has costs for individuals and societies due to behaviour restrictions and the cost of monitoring and enforcing the norm.

7.1 Individual evaluation

Norm evaluations made from an agent perspective are typically for two reasons: compliance and adoption. If a norm exists in a society then an agent must determine whether it will comply with that norm in the

situations when it is applicable. This is one type of norm evaluation, as the agent must decide whether compliance will bring it more value than violation. However, such an evaluation must take into account the enforcement of the norm—specifically, the probability of being punished and the extent of the likely punishment. A purely utilitarian approach would weigh up the expected net benefit from following the norm against the expected net benefit of violating it (including both the direct costs of possible punishment and the indirect costs of possible miscoordination with other agents complying with the norm). Rewards could include the benefits of improved coordination (if complying) or gaining an advantage over other agents (if violating). This utilitarian approach may be of limited use when evaluating a newly emergent norm that has no associated punishments (although things change if other agents have begun to punish violators, or reward compliance). Therefore, agents may use metrics other than pure utility. They may consider whether the new norm is coherent and consistent with their other norms, beliefs and goals. Agents may also consider the effect of compliance or non-compliance on their reputation amongst their peers and the trust that others have in them.

Agents evaluating strategies without regard to punishments are common in the literature as part of models investigating the spreading of norms. This can range from simple imitation of a single successful neighbour to an assessment of the local agent strategies. We examine a selection of these approaches in Section 8.3 where we discuss the role of influencer agents in norm spreading. These evaluations are purely concerned with the utility to a single agent; although they may result in a convergence to a common norm that benefits society, this is usually accidental.

Moving away from the purely utilitarian approach, Joseph *et al*. (2008) use the notion of coherence to evaluate a new norm for the purposes of deciding whether an agent should adopt a norm. In their approach, norms that are more coherent with existing norms and beliefs are evaluated more highly. Thagard's theory of coherence is used to measure coherence in terms of constraint satisfaction: associations between beliefs are seen as imposing either positive constraints, which reinforce those beliefs, or negative constraints, which weaken them. A coherent set of beliefs has few (or no) negative constraints between the beliefs, whereas an incoherent set of beliefs may have many negative constraints. Joseph *et al*. formalize this theory, and represent belief and norm sets as coherence graphs: nodes representing beliefs and weighted edges representing associations. When a new norm is proposed it is added to the graph, and the coherence of the combined graph can be assessed. Criado *et al*. (2010) present a norm compliance mechanism that also uses theories of coherence and consistency. In their approach, an agent evaluates a proposed norm with respect to its existing norms. If adding a new norm will make their norm set, or overall beliefs, incoherent or inconsistent then it is not accepted. Both of these approaches (Joseph *et al*., 2008; Criado *et al*., 2010) are notable for being purely individual strategies for evaluating a norm, in which an agent considers only the norm with respect to its existing set of norms. They require no assessment of whether the agent's peers consider the norm to be valuable. In contrast, Itaiwi *et al*. (2014) propose an agent level norm evaluation framework that combines utility, the degree of adoption of the norm by one's peers, and the consistency of the norm with respect to existing norms. They do not specify a mechanism to evaluate the utility or consistency (which they term the norm's morality), and calculating the degree of adoption requires a global perspective (although it would be possible to modify their approach to only consider the observable neighbourhood).

Andrighetto *et al*. (2010b) provide a model of norm *internalization* where an agent decides adopts the normative goals as their own and self-enforces the compliance with the norm. This process requires that an agent evaluates the norm in order to decide whether or not to internalize it. They implemented an internalization module in their EMIL framework that determines when to internalize a norm based upon two factors: norm salience and the cost of deliberating whether to comply with the norm. Salience is calculated from interactions with their neighbours, punishments observed and received, and from educative messages received from other agents regarding the norm. The cost of deliberation is a cumulative measurement of how long the agent has spent performing cost-benefit calculations concerning complying with the norm. If both salience and the cost of deliberation exceed a threshold then the norm is internalized.

Evaluating the norm in terms of personal utility or coherence and consistency with its existing beliefs can be valuable for an individual agent, but it does not necessarily shed light upon the value of the norm to society at large. An agent may hold selfish beliefs, or beliefs not shared by the majority of the society, and

in any case, simply because a norm is coherent and consistent with existing norms does not make it intrinsically of value.

The approaches in this section are useful for an agent seeking to determine whether or not a norm is beneficial to itself. However, norms are usually considered in a social context, in so far as they are shared rules, so it makes sense to consider their value in terms of groups of agents, rather than individuals. In the next section, we consider approaches that evaluate norms within a wider context—that of a society or computational system.

7.2 System evaluation

In this section, we review approaches to evaluate a norm from a system or societal perspective. For any such evaluation, it is necessary that the system or society have some set of goals that it wishes to achieve, so that the norm can be evaluated with respect to that set. However, there is no need for this set of goals to be imposed by a central authority, since the goals could be an aggregate of the individual user goals, or the valuation could be based upon a utilitarian ideal of the greatest happiness for the greatest number of individuals (Mill, 1863).

At the most basic level, evaluation can determine whether or not a norm will allow the system to fulfil its goals at all. Shoham and Tennenholtz (1995) take this approach in their study of the offline design of social laws, when they define a *useful law* as one that allows agents to perform their required tasks (represented as changing the environment from one state to another) without other agents interfering no matter what actions they perform, so long as all agents comply with the laws. For complex systems and social norms, this conception of usefulness is perhaps both too strong and too weak. It may be too strong, since a norm may help an agent reach its goals most of the time without absolutely ensuring that it will not occasionally fail due to interference from other agents (especially where agents may have conflicting goals); and it may be too weak because it allows no gradation of usefulness, and no notion of the fact that some norms may be more effective than others.

Morales *et al*. (2015) include a norm evaluation calculation in their norm synthesis mechanism (IRON). They evaluate a norm using two metrics: effectiveness and necessity. Effectiveness is a weighted measure of the ratio of successful norm fulfilments (instances where complying with the norm led to a desired state of affairs) to total fulfilments (including instances where complying led to an undesired state). So, for example, if a norm is complied with 10 times and leads to the desired state eight times, then the effectiveness is 0.8. Necessity is a weighted measure of the ratio of harmful violations (instances where violating the norm led to an undesired state of affairs) to total infringements (including instances where violating the norm had no ill effects). So, if a norm is violated 10 times and only two violations lead to a undesired state, then the norm has a necessity of 0.2. These measures are calculated over time using a reinforcement learning approach. This evaluation approach assumes that the results of all agent actions are observable and that compliance or violation of the norm can be explicitly connected to a desirable or undesirable state. In simple systems, this may be true, but in more complex ones complying with, or violating, a norm may have a more nuanced effect since norms may have indirect effects on agent behaviour beyond the success or failure of the interaction directly governed by the norm.

The notion of gradations of usefulness, or fitness, is key to evolutionary algorithmic approaches of norm design. In such approaches, a population of candidate norms is created and evaluated on a simulated system. The best performing candidates are reproduced, using evolutionary algorithm techniques, into the next generation and re-evaluated. This process continues until an adequately effective norm is found. For example, Bou *et al*. (2007) model a traffic scenario with norms controlling how vehicles should behave at road junctions (i.e. whether to give way to traffic coming from the right). The fitness was calculated based upon the number of factors, including the frequency of collisions and the speed of traffic flow. Genetic algorithms are used to increase the fitness of the norms in an iterative evolutionary process.

Simulation techniques have also been proposed to evaluate public policy in a variety of domains, including agriculture (Berger *et al*., 2006) and public utilities (Bunn & Oliveira, 2001). Dignum *et al*. (2009) examine the use of agent-based simulation to evaluate public policy in human societies and propose a simulation framework that encompasses three levels: first, individual agent personality and cognitive

Table 3 Summary of norm evaluation approaches

Approach	Norm adoption autonomy	Observability requirements
Joseph *et al*. (2008)	Required	None
Criado *et al*. (2010)	Required	None
Andrighetto *et al*. (2010b)	Required	Peers
Itaiwi *et al*. (2014)	Required	Peers, system
Shoham and Tennenholtz (1995)	Not required	System
Morales *et al*. (2015)	Not required	System
Bou *et al*. (2007)	Not required	System
Dignum *et al*. (2009)	Not required	System
Haynes *et al*. (2014)	Not required	System

traits; second, cultural aspects, such as existing social norms; and third, the macro-level effects upon society. Haynes *et al*. (2014) explicitly use a notion of the fitness for purpose of a norm in their work on estimating norm impact. Norm impact is defined as the difference caused by the existence of a norm upon the performance of an organization, with this performance being based upon the achievement of organizational goals. The organization is simulated both with and without the norm of interest in order to estimate the impact. The efficacy of such simulation techniques depends upon the accuracy of the model with respect to reality, or at least the parts of reality that relate to both the norm and the organizational, or societal, goals. Creating accurate simulations may be very hard, even with help from domain experts.

7.3 Analysis

The choice of norm evaluation approaches depend upon normative autonomy and observability. Table 3 summarizes the requirements of the norm evaluation approaches. The system evaluation approaches require that some entity in the system be able to observe the effects of the emergent norm across the entire system, so that this can be used to assess its value. The individual evaluation approaches (Joseph *et al*., 2008; Criado *et al*., 2010; Itaiwi *et al*., 2014) assume that the agents choose their own norms, and thus have norm adoption autonomy, whereas the other approaches make no such assumption. In fact, with the system evaluation approaches as presented, it would make little sense for an agent to choose their own norms, since they do not have the overall view of the system that is necessary for them to assess the value of a proposed norm. However, one could envisage a modified system approach where individual agents are presented with information about a norm's value (assessed at a system level), and are then given the choice whether or not to adopt it. We are not aware of any existing research that examines this possibility.

8 Mechanisms to encourage norm establishment

If a useful norm has begun to emerge in a multiagent society, but has not become established, it may be useful to encourage more agents to comply with the norm so that the society converges upon it as a solution to whatever social problem it helps. This is sometimes referred to as the norm spreading through society, and the process may be as gradual or organic as that implies. Social norms may spread as agents observe their neighbours' behaviour and copy that which appears useful, or agents may communicate with others and persuade them to behave like they themselves behave. However, entities within the society may also impose mechanisms to encourage the rapid adoption of a norm they deem to be beneficial (either for themselves or society in general). Also, both organic spreading and imposed mechanisms may be acting to increase norm adoption at the same time, and may interact.

As an example of encouraging the establishment of a beneficial norm, consider the norm obliging a car driver to wear a seat belt in the United Kingdom. The benefit of wearing a seat belt restraint was recognized as early as the 1930s, and the three-point safety belt common today was invented in 1958, however, the initial adoption was relatively slow. In 1967, a law made it compulsory to fit new cars with seat belts in the

United Kingdom. This legislation made it easier for drivers to comply with the norm, since they did not have to have a seat belt fitted after purchase. In the 1970s, the UK government began a series of campaigns using popular public figures to educate drivers of the benefits of complying with the norm, however, compliance rates did not rise above 40%. In 1983, the government introduced legislation that mandated a punishment (a fine) for non-compliance, along with an intensive advertising campaign to inform people of the law and the penalties for violation. Studies showed that the compliance rate was ~40% a month before the law came into effect, 50% on the day before and 95% the day after it came into effect (Broughton, 1990). Compliance has subsequently remained around 95%.

The seat belt example illustrates several of the mechanisms used for encouraging norm establishment: educating the agents about the benefit of the norm, using influential members of the society to promote the norm and introducing punishment for non-compliance. It also demonstrates how a society can formalize a beneficial norm by creating a legal norm with a specified punishment (a monetary fine) and agents responsible for monitoring compliance (traffic police). In this section, we discuss these and other mechanisms in the context of multiagent systems. Specifically, we review research on the following: the role of metanorms (norms about norms) in norm establishment, mechanisms to control how norms are introduced and changed, the use of influencer agents to promote norm compliance, the effect of incentivization on norm compliance, the use of rewiring mechanisms to alter network topology and encourage norm spreading, and detecting and resolving normative conflicts.

8.1 Metanorms

Metanorms are norms about norms, or in other words, rules that concern the usage of norms in a society. With regards to norm emergence, research on metanorms has concentrated on a rule that obliges agents to punish norm violators. Since delivering punishment usually has a cost, agents may prefer not to punish norm violations that they observe. The effect of the metanorm is to encourage agents to punish violators, since not punishing them may itself bring about punishment. Axelrod (1986) makes use of metanorms in his norms game in order to bring about norm emergence. In his norms game, each agent chooses to either cooperate (comply with a norm), or defect (violate it). Defectors may be observed by other agents with a known probability, who then have the chance to punish the defection. Each agent has two properties: vengefulness and boldness. If an agent's boldness is greater than the probability of being observed it defects and receives a temptation payoff and the other agents suffer a loss. If another agent observes the defection it chooses to punish with a probability of its vengefulness property. A punished agent receives a large negative payoff, but the punishing agent also suffers a loss to reflect the cost of enforcement. Axelrod's norms game uses an evolutionary approach: agents with a high utility score at the end of each round are reproduced twice into the next generation to play the next round of the game, and there is a small chance of a random mutation changing the agent properties.

Norm establishment is represented by a population with high levels of vengefulness and low levels of boldness since this reflects a society that punishes defectors while mostly complying with the norm. Axelrod shows that the basic norms game rarely results in the establishment of a norm, due to the cost of enforcement effectively punishing those with high levels of vengefulness. He therefore proposes the use of a metanorm, whereby agents are themselves liable to be punished if they are observed allowing violations to go unpunished. His experiments using the metanorm result in the establishment of the norm in most cases.

Subsequent investigation into Axelrod's norms game (Galán & Izquierdo, 2005; Mahmoud *et al.*, 2010) has found that extending the number of generations lead to the eventual collapse of the norm, due to the chance of a sequence of unfavourable mutations taking advantage of the evolutionary nature of the agent reproduction. However, this weakness in Axelrod's original game does not invalidate the use of metanorms to aid norm emergence. Mahmoud *et al.* (2015a) show that metanorms can be used in a more realistic context to promote the emergence of norms. Specifically, they investigate a model that uses learning, rather than evolution, to change the agent properties, and where agents only punish according to the defections that they observe, in contrast to Axelrod's game where agents punish non-punishers even if the initial violation is not observed.

8.2 Controlling norm introduction and change

As well as metanorms concerned with the *usage* of norms, multiagent systems may also have rules and mechanisms for changing and introducing norms. These are analogous to Hart's (2012) rules of recognition and change for laws in human societies (see Section 3). The nature of these rules, if they exist, determines how a norm emerges and becomes established in an agent society. Therefore, in this section we consider different approaches that have been proposed to control how norms may be introduced and changed in multiagent systems.

Artikis (2012) presents an infrastructure to allow the runtime modification of system norms by agents within the system. The infrastructure uses meta-protocols that specify how and when norms may be changed (e.g. which agents may propose new norms), and also higher levels of meta-meta-protocols that determine how those meta-protocols may themselves be changed.

In order to evaluate proposed norm modification, Artikis suggests two factors: the similarity of the modified norm to the original norm, and its expected utility (which is domain specific). To derive the similarity, they model the norm specification as a metric space (Bryant, 1985) whereby the 'distance' between the existing norm and the proposed modified norm can be calculated using some function. In this way, the system designer can constrain norm changes so that only gradual, incremental changes (represented by a small distance) can be made. As an alternative, in previous work, Artikis (2009) proposed limiting the maximum distance of new norms from some 'desired' norm specification in order to allow the system designer more control.

Another agent-based norm modification approach is proposed by Riveret *et al*. (2014). In their approach, agents generate possible norms by considering their own experiences: for example, by learning what behaviour gives the greatest reward in a certain situation. They then each propose a possible norm for consideration by the other agents in the system. The most common proposal is voted on by all agents and, if a majority support it, it becomes a new system norm.

Tinnemeier *et al*. (2010) present a rule-based set of programming constructs to facilitate the operationalization of norm introduction and change during runtime. Their approach uses the concepts of norm schemes and norm instances. Norm schemes are the underlying rules of conduct for the system designed to further the design goals of the systems, and norm instances are the obligations and prohibitions concerning specific system states. Separate change rules may be specified for schemes and instances to allow fine-grain control: for example, agents may be able to change norm instances to reflect their localized knowledge, but not change the norm schemes which may require a more global perspective. While they do not specifically address emergent norms, their approach could be used to implement other techniques for engineering emergent norms.

8.3 Influencers

The metanorm approaches assume that peers enforce norms, and that no agents are privileged over another. However, it may be natural and desirable that some agents either have more power to enforce norms than others (because they are empowered to do so by the system), or that some agents are willing to expend more effort persuading others to follow norms. These agents are known as *influencers*, and the use to aid the spreading or emergence of a desired norm has been the subject of research. Such agents try to influence the behaviour of other agents by complying with the norm and by punishing violators or rewarding compliers, and are themselves invariant in their adoption of the norms—that is, an influencer will not cease to punish or reward. An example in human societies are police officers: they (should) punish law-breakers consistently. Norm entrepreneurs (or norm innovators) are agents who devise a new norm and try to convince others to follow the norm, either by persuasion or by example (Finnemore & Sikkink, 1998). Human history shows many examples of individual or small groups of norm entrepreneurs, that successfully effected changes to the prevailing norms of the time. For example, the abolitionists who led to the end of the slave trade, and the suffragettes who helped to bring about women's suffrage, were norm entrepreneurs.

Sen and Airiau (2007) investigate the spread of norms via social learning where each agent learns behaviour from interacting with other agents. As part of their investigation they examined the effect of *fixed agents*, who do not learn, but always stay with the same strategy. In a simple coordination game

where agents have a binary choice of behaviour they find that using as few as four fixed agents within a population of 3000 almost always leads to a convergence with the fixed strategy. These fixed agents can be seen as limited influencers, in so far as they persuade only by their actions and the other agents converge in order to accommodate these immovable objects in their society. However, the model is very simple, with only two choices of strategy and no notion of punishments (beyond the cost of using a defecting strategy). Hoffmann (2005) explores the effect of norm entrepreneurs on norm emergence in agent-based simulations. In this model, agents have three candidate norms (out of a possible seven) at any one time and attempt to converge to a common group norm. The norms are scored according to how close they are to the group norm. The norm entrepreneurs suggest norms to the other agents, who use the suggestion to replace their lowest scored norm. It was found that entrepreneurs aided the establishment of a common norm.

Franks *et al*. (2013) examine the use of influencer agents to affect the convergence of a convention in a more sophisticated model. They use a linguistic domain where agents build up lexicons that map words to concepts. Agents start with randomized lexicons and over a series of rounds, they try to communicate with each other. This communication is naturally more successful if agents have more similar lexicons. The quality of this communication is measured and each round they share their lexicons (or parts of them) with other agents, and update their own lexicons based upon this level of quality (so high-quality lexicons are propagated). Over time, the agents may converge on a common lexicon, which results in more successful communication for the whole society. This domain provides many possible strategies to converge upon. Influencer agents are represented by agents with the same fixed high-quality lexicons. It is found that a relatively small number of influencers can aid the convergence of a convention. Network topology affects the results: when the influencer agents are well-connected in a communication network fewer of them are needed. As well as fixed strategy agents, Franks *et al*. examine the effect of non-fixed agents starting with complete, high-quality lexicons. These non-fixed agents could alter their strategies. They were also found to improve convergence, though not to the extent of the fixed strategy agents. This is perhaps analogous to highly skilled agents prepared to teach by example, but not as single-minded as norm entrepreneurs. Finally, flawed influencer agents were investigated: fixed strategy agents with poor-quality lexicons. These flawed influencers were found to have a detrimental effect on convergence, with few such agents required to prevent a useful lexicon arising in the society.

Savarimuthu *et al*. (2008) propose a mechanism to spread norms through a society of agents by the means of role model agents. Agents choose the most successful agents in their neighbourhood as role models, and these role models pass advice (in the form of a norm) to their followers. The follower agents then modify their norm based upon their role model's advice. This is similar to learning by observation and imitation of successful agents, but because the communication is explicit the observer need not infer the successful strategy. However, it does open up the risk of malicious role models exploiting their followers for personal gain.

As well as individual agents, or a class of agents, influencing their peers, it is possible to have an approach where there is a system-wide mechanism for recommending norms to agents. Savarimuthu *et al*. (2013b) present an architecture for such a norm recommendation service that has four distinct phases: identification, classification by salience, classification by stage in the norm life-cycle, and the recommendation itself. The identification is performed using the detection of sanctioning behaviour (as in Savarimuthu *et al*. (2013a) described in Section 6). Salience is a measure of how important the norm is to the system and is determined by the frequency of the action targeted by the norm and the probability of a norm violation being sanctioned. For example, high frequency actions that are frequently punished are deemed to be very highly salient. The life-cycle is determined by the change of salience over time: for example, newly detected norms are deemed to be emergent, norms that have risen above a specified salience threshold are deemed to be mature. The determination whether to recommend a norm is based on system-specific heuristics using the salience and life-cycle. While Savarimuthu *et al*. suggest that the recommendation architecture could be used by individual agents, its reliance upon observations perhaps makes it more suitable as a system-wide service able to observe globally (or at least more widely than a single agent). It could clearly be used in concert with system-controlled influencer agents such as those detailed earlier in this section.

8.4 Incentivization

In this section, we examine the use of *extrinsic incentives* to encourage agents to comply with norms. Incentives can include rewarding desirable behaviour, punishing undesirable behaviour or both. Complying with a norm may bring about intrinsic benefits to an agent due to improved coordination or cooperation with neighbours, and violating a norm may likewise cause intrinsic loss of utility. For example, if an agent chooses to drive on the wrong side of the road it will certainly suffer delays and probably collisions. On the other hand, if it complies with the driving lane norm, it reaps the benefit of improved coordination and can travel smoothly with the rest of the traffic. While these intrinsic incentives may be enough to encourage desired behaviour, this is not always the case. The undesired behaviour may be individually rewarding for an agent, while bad for the society as a whole (e.g. in a tragedy of the commons situation), or an agent may be unable to calculate the intrinsic benefits for itself.

Therefore, on top of these intrinsic incentives, societies may need to impose extrinsic incentives to encourage (or discourage) norm formation and adoption. In the driving example, being caught violating the driving lane norm will result in legal penalties imposed by society. The driving laws, and their enforcement, both increase the potential cost of behaving badly, and make explicit that cost, as opposed to the uncertain cost of delays and collision.

The nature of the incentive must depend upon both the deliverer and the target of the incentive. With purely peer-based incentivization where agents have no direct power over each other, incentives are limited to the offering of rewards for good behaviour, or punishment via the withdrawal of services or cooperation. For example, agents could shun norm violators and refuse to deal with them. The use of trust and reputation systems can facilitate peer-based incentivization, since they allow a long-term view of the trustworthiness of agents and, with respect to reputation, allow agents to share their opinion of other agents (Castelfranchi & Falcone, 1998; Castelfranchi *et al*., 1998). Where the incentivization is delivered by the system itself, there are other options—violators could be barred from services, or even ejected from the system.

With respect to humans, there is some evidence that the use of extrinsic rewards can reduce the intrinsic motivation to perform desirable behaviour (Gneezy *et al*., 2011), especially if the rewards are later removed. If behaviour comes to be performed purely to gain reward (or avoid punishment) then the salience of intrinsic rewards may be reduced. For example, children encouraged to read by monetary rewards may come to see reading as a chore (albeit a profitable one), and may not read unless they are offered a reward.

With computational agents, rather than humans, the effect of incentivization is simpler and more easily studied. However, when dealing with autonomous agents with uncertain motivations, whose internal reasoning is opaque, it can still be challenging to design an appropriate incentivization systems. In general, there are two questions that must be answered to determine whether an incentive system is appropriate: is the correct behaviour rewarded (or punished), and is the degree of reward (or punishment) sufficient to modify the behaviour?

Extrinsic incentives may be delivered by peers, where any agent has the ability to punish or reward behaviour, or by specific agents empowered by the system to administer punishment or deliver rewards. In this section we are concerned with incentivization administered by the society or system, either indirectly, through mechanisms which control how peers may punish or reward others, or directly, via centralized mechanisms or specified agents. Peer incentivization has been studied in relation to metanorms (described in Section 8.1) for the purpose of encouraging agents to punish norm violators, however, it may also be useful to restrict peer punishment.

While punishment has a cost which agents may be unwilling to pay (leading to under-punishment), agents may also over-punish: either to harm rivals, or to impose their view of the severity of a norm violation upon society regardless of the opinions of other agents. Over-punishment can be wasteful of resources (if the cost of a punishment is proportional to its severity), or can be harmful in other ways. For example, in an open system such as a market place, where accidental norm violation is possible, a fear of over-punishment can deter agents from joining the market and so reduce the potential customers. For these reasons, the administering of punishments can itself be a source of asocial behaviour. One mechanism to

prevent over-punishment is a consensual institution of peer punishment (Casari & Luini, 2009), where violations are only punished when more than one agent agrees with the punishment. This prevents single agents from punishing others for purely individual reasons. An alternative approach is proposed by Faillo *et al.* (2013), in the context of punishing free-riders in a cooperative scenario. In their approach, agents can only punish those agents who contribute less than they themselves contribute. This prevents antisocial behaviour where low contributors punish high contributors, either as retribution for previous sanctions, or simply out of spite. Both these approaches seek to ensure that a punishment is applied in appropriate situations, rather than in an arbitrary fashion. The other factor to consider is whether the degree of incentivization is appropriate, or, in other words, is there enough of an incentive to ensure that behaviour is modified while also ensuring that not too many resources are expended?

A static punishment system, where each violator receives the same sanction for each violation has the advantage of simplicity, but it suffers from the problem that agents may react differently to the same degree of punishment. It also punishes first-time offenders the same as perpetual recidivists. Human legal systems often vary the punishment for crimes based upon the past criminal history of the offenders, with repeated law breaking being dealt with far more harshly than a first-time offence, or monetary fines scaled to match the wealth of the offender. In the context of multiagent systems, the notion of dynamic punishment, where the punishment for norm violation changes in response to agent reaction to that punishment, has been studied both with a simple escalating model (Miceli & Bucci, 2005; Emons, 2007) where repeated violations incur a greater punishment, and using machine learning techniques to refine incentives depending upon the reaction of agents (Villatoro *et al.*, 2011a; Mahmoud *et al.*, 2012b, 2015b).

Increasing the punishment for an agent which violates a norm repeatedly, is analogous to the custom in some human legal systems for increasing sanctions for recidivists. The concept is based upon the assumption that the initial sanction did not make the violation costly enough to the agent (i.e. they still gained enough benefit from the violation to make it worth their while to repeat the violation), or that the first violation may have been a mistake (in which case a relatively small sanction may be enough to ensure that they take more care in the future). Emons (2007) examines the effect of varying sanctions on agents. In his scenario, agents can violate norms by accident or purposefully twice in order to gain utility from the violation; the scenario was intended to represent a commitment to criminal behaviour rather than a one-off intentional violation. If detected, violation was punished by fines that reduced the agents' utility; two strategies were investigated: a small fine for an initial offence, followed by a larger fine for a second offence, or vice versa. It was found that if violation gave a large benefit, then a large fine for the second offence was optimal, since it made honesty more attractive. Bearing in mind that accidental violation was possible, a large fine for a first offence serves to make honesty less attractive, since the punishment for a second offence would be lower and agents who accidentally violate a norm have little to lose from violating it again. On the other hand, if the benefit of violation is low then a large fine for a first offence is optimal since it makes criminality less attractive—since detection is not automatic, a persistent violator is more likely to be detected only once than twice. As mentioned above, in this scenario a deliberate intention to violate a norm requires the agent to violate it twice to represent a commitment to a life of crime.

Norm violators may suffer incidental costs as well as direct sanctions. In particular, they may have reduced prospects of legitimate income in the future. For example, humans convicted of criminal behaviour may find it harder to find legal employment, and computational agents working within systems with trust and reputation mechanisms may find that past norm violations restrict their options. This reduction in future income serves as a deterrent to first-time violators but can reduce deterrence in repeat offenders, since compliance brings about lower rewards. Miceli and Bucci (2005) show that, in such a situation, an escalating punishment mechanism is optimal, since it makes continual violation less and less beneficial to the agent.

While increasing or decreasing punishments based upon repeated violations can be effective, dynamic punishment has been proposed as a way of altering the degree of punishment in response to circumstances in a more flexible way. With dynamic punishments, the degree can either go up or down as required, so that agents are neither over nor under punished.

Villatoro *et al.* (2011a) propose a dynamic adaptation heuristic that varies punishment based on the number of defectors (agents violating the norm) in the society. If the number of defectors is increasing and

above a set tolerance threshold, then the punishment is increased by a small amount. If the number is decreasing or below the tolerance threshold, then the punishment is decreased by a small amount. Using the dynamic punishment reduced the amount of punishment that was required to establish a norm in their model (a variation of the Prisoner's Dilemma game). Varying punishment based on the number of defectors has two restrictions: first, the entity performing the punishment must either know the number of defectors, or be regularly informed of the correct amount of punishment to inflict, and; second, the punishment amount is determined based on the behaviour of the entire population and not the individual behaviour of the agents. If the total number of defectors is low, then a consistent defector will always receive a low degree of punishment. An alternative approach is to vary punishment dynamically based on the individual behaviour of agents. Mahmoud *et al.* (2012b) propose such a punishment mechanism in order to bring about norm establishment in the context of Axelrod's (1986) norms game (described in Section 8.1). In their mechanism, each agent keeps track of past interactions and modifies how they punish violators based upon the number of previous violations. Only a specified number of the most recent interactions are considered; older violations are considered forgiven. This approach was found to bring about norm establishment with less punishment than a static punishment approach.

Mahmoud *et al.* (2014) propose an information-based incentivization framework, capable of supporting monetary rewards with any number of other relevant incentivization mechanisms, offered to an entity in the form of informational incentives. This framework allows learning and reasoning about the effect of various incentives on an entity's behaviour (which is not known in advance and may change over time), and reflecting such reasoning on the design of more effective (personalized) future incentives for the entity.

8.5 Rewiring networks

Norms can spread between agents as they interact and change their behaviour to accommodate the behaviour of others, or learn successful strategies from their neighbours (e.g. via influencers as described in Section 8.3). The agents and the connections between them can be represented as a network or graph, with the agents being nodes of the graph and the interaction links being the edges. The topology of the connections can affect how norms spread through an agent population (Villatoro *et al*., 2009; Sen and Sen, 2010). In particular, norms have been found to spread more quickly in networks where agents interact with many neighbours, as opposed to networks where agents have fewer neighbours. Some work has examined the impact of different topologies on norm establishment. For example, Savarimuthu *et al*. (2007) consider the *ultimatum game* in the context of a role model that provides advice on whether to change norms in order to enhance performance, and provide experimental results for random and scale-free networks. Delgado *et al*. (2003) study norm emergence in coordination games in scale-free networks, and Sen and Sen (2010) similarly examine rings and scale-free networks in a related context. Additionally, Villatoro *et al*. (2009) explore norm emergence within lattices and scale-free networks. Mahmoud *et al*. (2013) propose dynamic policy adaptation to aid the establishment of norms using metanorms in scale-free networks. While these efforts provide valuable and useful results, the context of application has tended to be limited, with only two agents involved in a single interaction, rather than a larger population. This simplifies the problem compared to those in which *multiple* agents involved in a single interaction can impact on norm establishment. For example, norms may emerge in group situations such as within newly created human teams solving a group task (Bettenhausen & Murnighan, 1985) or in collaborative enterprises such as open source software projects (Savarimuthu & Dam, 2013) where developers work on the same task over a period of time, but do not necessarily interact directly.

Since the topology of the agent network connections influences the spread of norms, rewiring those connections has been proposed as a method of encouraging norm emergence. This rewiring can be purely *ad hoc* as individual agents choose which of their neighbours to interact with based upon their behaviour (Zimmermann & Eguíluz, 2005; Zhang & Leezer, 2009; Griffiths & Luck, 2010), or it can be performed with the specific aim of encouraging or discouraging norm establishment (Garlick & Chli, 2009; Villatoro *et al*., 2011b).

Zhang and Leezer (2009) examine the effect of selfish *ad hoc* rewiring upon the emergence and spread of norms. In their approach, each agent seeks to maximize its utility without concern for others and rewires

its own connections in order to do so. Unrewarding connections are broken and rewarding ones maintained. They evaluate their approach with agents playing three games (Prisoners Dilemma, Stag Hunt and Pure Coordination) and found that this selfish rewiring led to a faster convergence of behaviour, and so establishment of a social norm, than simply learning from interactions in a static network. Their agents do not consider the experiences of other agents, nor do they observe the behaviour of others, so it is a purely individual approach that does not rely on trust in other agents.

If agents can observe the interactions of their neighbours, this information can be used to supplement their own experiences. Zimmermann and Eguíluz (2005) use this approach to examine the effect of allowing agents to remove connections with neighbours within an iterated Prisoner's Dilemma game. Removed neighbours are replaced with a random agent, so agents have no control over their new neighbours. Agents modify their strategy based upon the most successful of their neighbours, and these successful agents do not change their neighbours. The agents typically converge into either fully defective or fully cooperative societies, with the cooperative societies dependent upon successful cooperative 'leader' agents arising with stable local networks (or chains) of cooperative followers.

As well as observing the interactions of neighbours, agents may also be able to modify who they interact with based upon the reported experiences of those neighbours. This is, in effect, a reputation system where norm violators gain a bad reputation which leads to other agents shunning them. Griffiths and Luck (2010) propose such a rewiring mechanism. They focus upon agents using rewiring in order to punish defectors: essentially removing norm violators from their local neighbourhood networks. Agents do not use only their own experiences, but also those of their neighbours. Although they do not explicitly address emerging norms, this punitive rewiring could be used to reinforce norms that are not fully established and so hasten convergence, and the use of information from neighbours allows a social consensus to emerge about what constitutes defection from a norm.

The above mechanisms facilitate convergence towards a norm indirectly—agents copying successful behaviour or shunning defectors are seeking to maximize their own rewards, rather than explicitly trying to bring about the establishment of a norm. However, it is also possible to use rewiring in a more conscious attempt to influence the spread of a norm. The use of restricting communication links to discourage unwanted behaviour is examined by Garlick and Chli (2009). They use an agent-based model of social unrest, where agents' behaviour is influenced by those they interact with, and investigate the effect of implementing curfews whereby agents are prevented from interacting with their neighbours (unless they take a risk of arrest by breaking the curfew). Unsurprisingly, the results show that such curfews inhibit the spread of unrest.

A more sophisticated approach is proposed by Villatoro *et al.* (2011b). Their method focusses on removing metastable subconventions that otherwise prevent the widespread establishment of a norm. Subconventions are regional norms adopted by subsets of agents within the society, perhaps due to local network topology or local agent preference. For example, if a subset of agents are relatively isolated from the main society then a subconvention could be locally reinforced and not allow another (potentially more useful) norm to spread. They propose the use of observation and rewiring (which they collectively term *social instruments*) to facilitate norm convergence in an agent society. In their approach, agents observe interactions outside of their immediate neighbourhood and learn from those interactions. By observing widely, agents can modify their behaviour based upon global, rather than purely local, norms and so subconventions can be dissolved. Their second social instrument is rewiring, where agents change their links to other agents to overcome topological bottlenecks that are restricting the spread of norms. Agents identify frontier regions, where two conflicting norms meet, and further determine if a *self-reinforcing structure* (SRS) is present (where the topology maintains a local metastable subconvention). Figure 3 shows examples of the two types of SRS that their approach looks for—the caterpillar, and the claw. They are typified by the presence of 'hangers' that only connect to the SRS (or other SRSs), and not the wider world. For example, in the claw structure (Figure 3(b)) nodes C, D, G and H are hangers, and the structure consisting of B, E and F is another claw. If such a structure is found, the agents rewire their local network connections in order to remove the SRS and facilitate the dissolution of the subconvention.

Mungovan *et al.* (2011) also propose a method to remove local self-reinforcement of conventions. They introduced the idea of weighted random interaction by which agents are able to interact with random

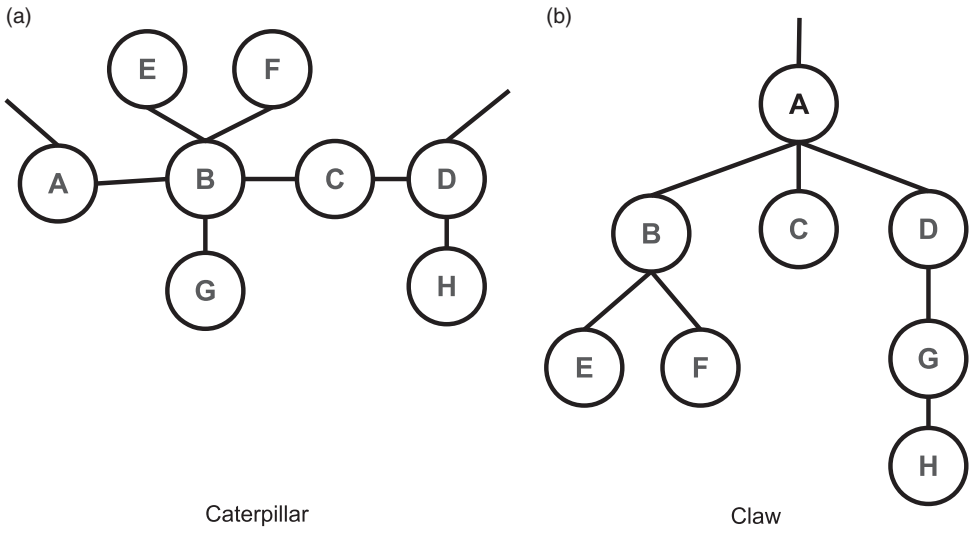

Figure 3 Self-reinforcing structures (after Villatoro *et al.*, 2011b); (a) caterpillar; (b) claw

members of the population based on the distance, so the closer an agent is to another, the more likely there will be an interaction between these two agents. Their results suggest that dynamic interaction helps in easing emergence especially in breaking local biases that are normally hard to break.

The research on rewiring agent connections shows that it can be an effective way of encouraging the establishment of a norm and removing conflicting local norms sustained by topological reinforcement. It can also be used as a sanctioning and enforcement mechanism once a norm is established.

8.6 Normative conflict detection and resolution

In the previous section, we describe methods to remove local subconventions by means of rewiring networks. This is an example of resolving conflicts between norms so that one can become fully established across the society. A normative conflict occurs when one norm obliges behaviour that is prohibited by another norm, or when two norms oblige actions that cannot both be performed: for example, obligations to attend two social events at the same time (Elhag *et al.*, 2000). Such conflicts may prevent norms emerging in a society because agents cannot comply without violating other norms. For this reason, detecting and resolving such conflicts can aid the emergence of norms. Additionally, the resolution technique can be tailored to favour the norms preferred by the system designers. The research on normative conflict detection and resolution is extensive, so we do not attempt to fully survey the field. Much of the research focusses on individual agent assessment of the value of the norm in an attempt to choose the best one. For example, Criado *et al.* (2010) (see Section 7) that examines whether a new norm is consistent with an agent's existing norms, or dos Santos Neto *et al.* (2012) in which an agent weighs up the benefits and costs of complying with each possible norm and chooses the most individually beneficial norm). We focus on system-based approaches, since if each agent resolves normative conflict in its own way they may disagree over which norm to favour and this will not aid in the encouragement or discouragement of an emerging norm.

Vasconcelos *et al.* (2009) detail an approach to detect and resolve conflicts between norms that regulate agent actions. They represent norms as constraints over variables and conflict occurs when the constraint variables of a prohibition overlap with those of an obligation or permission. This is detected, during runtime, using first-order unification. If a conflict is detected, it is resolved by curtailing one of the norms so that its scope is reduced—that is, it no longer overlaps with the other norm. Which norm to curtail is determined by system policies. The work suggests two policies, based on the legal principles of *lex posterior* (where the older norm is curtailed) and *lex superior* (where the norm proposed by the weaker authority is curtailed). A similar approach of unification and curtailment is taken by Vasconcelos *et al.* (2007).

Sensoy *et al.* (2012) propose a language for representing norms, OWL-POLAR, based upon the Web ontology language OWL-DL. They provide an algorithm that detects possible conflicts by creating

Table 4 Summary of approaches encouraging or discouraging norm establishment

Approach	References	Norm autonomy	Observability requirements	Enforcement
Metanorms	Axelrod (1986)	–	Peers	Peers
Metanorms	Mahmoud *et al.* (2012b)	–	Peers	Peers
Influencers	Sen and Airiau (2007)	–	Peers	–
Influencers	Hoffmann (2005)	–	Peers	–
Influencers	Franks *et al.* (2013)	–	Peers	–
Influencers	Savarimuthu *et al.* (2008)	–	Peers	–
Incentives	Casari and Luini (2009)	–	Peers	Restricted peers
Incentives	Faillo *et al.* (2013)	–	Peers	Restricted peers
Incentives	Emons (2007)	No	System	System
Incentives	Miceli and Bucci (2005)	No	System	System
Incentives	Villatoro *et al.* (2011a)	No	System	Restricted peers
Incentives	Mahmoud *et al.* (2012b)	–	Peers	Restricted peers
Rewiring	Zhang and Leezer (2009)	–	Self	Peers
Rewiring	Zimmermann and Eguíluz (2005)	–	Peers	Peers
Rewiring	Griffiths and Luck (2010)	–	Peers	Peers
Rewiring	Garlick and Chli (2009)	No	System	System
Rewiring	Villatoro *et al.* (2011b)	No	Extended peers	–
Conflict resolution	Vasconcelos *et al.* (2009)	No	System	–
Conflict resolution	Sensoy *et al.* (2012)	No	System	–
Conflict resolution	Günay and Yolum (2013)	No	System	–

canonical states of the world where each norm is complied with and using an ontology consistency checker (such as Pellet (Sirin *et al.*, 2007)) to see if such compliance is possible for all norms. Resolution is performed by one norm overriding another, either by *lex posterior* or *lex superior* (as in Vasconcelos *et al.*, 2009), or *lex specialis*, where a more specialized norm overrides a more general form[8]. This latter is possible due to an algorithm they detail that determines whether one norm can be subsumed by another.

Constraint satisfaction techniques can be used to detect conflicts between norms. Günay and Yolum (2013) treat obligations as constraints on agent behaviour and represent a set of obligations as a constraint satisfaction problem. Such a problem can be run through a problem solver to see if all the constraints can be satisfied. If not, then there is a conflict that must be resolved.

8.7 Analysis

Table 4 summarizes the approaches used to encourage or discourage norm establishment. The enforcement method and observability requirements are most important, since many of the approaches rely on peer pressure to spread or discourage norms: either direct sanctions (such as with the metanorm approaches) or via indirect measures (such as rewiring to exclude norm violators). The assumption of norm adoption autonomy is irrelevant to most of the approaches, since they do not specify where the norms come from, but some of the system level approaches do assume that a norm has been mandated by a controlling authority, or mandate how and when agents must punish violators.

Most of the research effort on norm establishment has been focussed on peer-based systems, presumably since more centralized controlled systems can often establish a norm by fiat. However, this means there is a gap with respect to systems that seek to exert some degree of control over its norms, while being unable to impose those norms directly: for example, systems that are open to autonomous, heterogeneous agents that wish to maintain specific community standards.

[8] For example, if there is a norm obliging vehicles to travel under 70 miles/hour, and a norm obliging trucks to travel under 50 miles/hour, the latter is a more specialized norm since a truck is a type of vehicle.

9 Conclusions and open issues

In this paper, we have examined the concept of engineering emergent norms for the benefit of a multiagent system. We have identified the three main steps and reviewed the literature concerning each step. These three main steps are: first, the emergent norm must be identified; second, the norm must be evaluated to determine if it is useful; third, norm spreading must either be encouraged or discouraged, depending upon its utility. A summary of the steps and the papers reviewed is shown in Figure 4.

While we have examined each step separately, some researchers present work that integrates several steps. In Table 5, we list and compare these approaches.

In a society of autonomous agents, as in any complex system, emergent behaviour is extremely likely. Indeed, it has been suggested that the manifestation of emergent behaviour is a property of *all* complex adaptive systems (Holland, 1992). This behaviour may manifest as emergent social norms as agents interact and try to solve coordination and cooperation problems. How that emergent behaviour is handled may affect both individual members and the society as a whole. Allowing emergent norms to spread naturally is the simplest approach, but raises a number of risks: unhelpful norms may emerge, helpful norms may spread very slowly and unevenly, and rival norms may emerge and lead to normative conflict. To mitigate these risks it may be desirable to *engineer* the emergence of norms: first, to encourage the swift spread of useful norms throughout the society; second, to discourage harmful norms; and, finally, to ameliorate normative conflicts.

The engineering of norms can be attempted by individual agents, via leadership or norm entrepreneurship, by groups of agents, by the society as a whole, or by an entity outside of the society (e.g. if the agent society is designed by humans to perform a task).

The identification of explicit norms has been studied from an agent perspective (Savarimuthu *et al*., 2010, 2013a; Oren & Meneguzzi, 2013; Cranefield *et al*., 2015), and the identification of emergent properties has been studied from a system perspective (Chen *et al*., 2007, 2009). Implicit norm identification is a feature of a number of norm spreading and emergence models, where agents observe their neighbours and look for successful strategies, or common behaviour.

To our knowledge there has been no research into group identification of explicit norms, where sets of agents combine their knowledge in order to identify emergent behaviour in their society. At a basic level this could be performed by a simple aggregation of individual beliefs about the existence of norms. However, for emergent behaviour, rather than existing norms, this may not be sufficient since the behaviour may itself be only partially observable by each individual agent. In other words, the true macro-level behaviour may only be recognized by aggregating observations rather than beliefs about norms. This is particularly relevant in the case of emergent interlocking norms, where norms depend upon, and interact with, other norms (y López & Luck, 2004). It may be the case that one group of agents identify a new norm and another group identify a separate but interlocking norm: only by combining their knowledge could a full understanding be gained.

The work on detecting emergent behaviour from a system perspective, such as the event-based work of Chen *et al*. (2009), and rule-mining (Van der Aalst *et al*., 2003) does not directly address the detection of norms and conventions. In our opinion, it would be valuable to place this work within a formal model of norms, so that concepts used in norm research, such as degree of emergence, salience and maturity can incorporated into these detection methods.

From an agent perspective, norm evaluation has been studied with respect to compliance with, or adoption of, a norm. The former, reasoning about compliance, is typically performed purely on a case-by-case basis, with the agent deciding whether it is useful to comply with a norm in a particular context. In contrast, the latter is usually treated as an all-or-nothing affair, with the agent judging whether the goals of the norm are compatible with its other goals, or, if not, whether the normative goals are more important. If a norm is adopted then the agent adopts the normative goals and, if necessary, adjusts its other goals. From a societal, or system, perspective, the value of a norm has been studied with respect to how well it helps a society achieve its overall goals. However, current approaches are either tightly focussed on individual interactions, or rely on omniscience and clear societal goals. Also, such approaches are centralized.

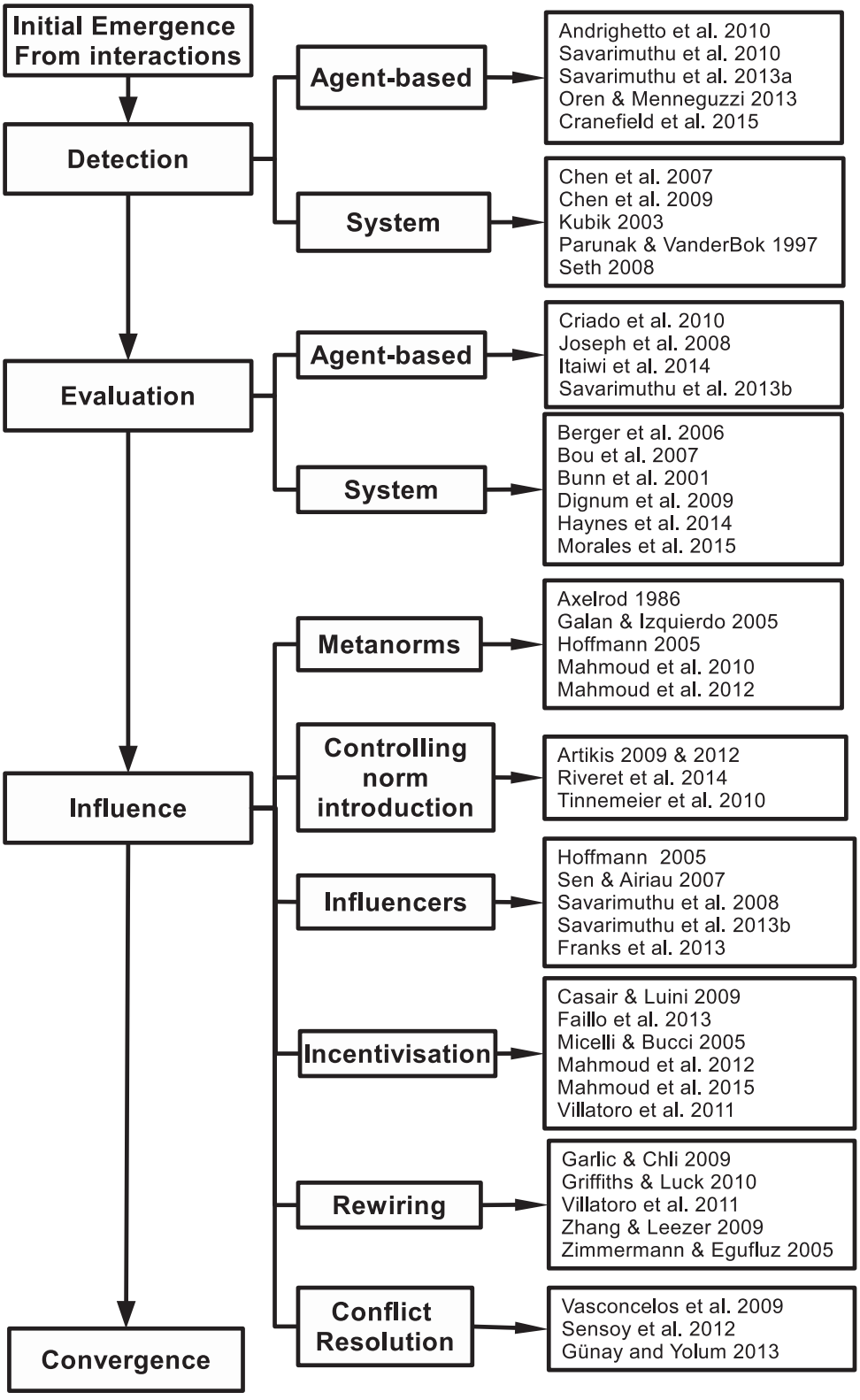

Figure 4 The main steps of engineering emergent norms, and related research

Therefore, there is a need for mechanisms where groups of agents can mutually decide on the value of a norm in a distributed fashion. Such mechanisms would need to overcome the typical issues of distributed decision making: individuals may not have the skills and knowledge to evaluate

Table 5 Summary of approaches spanning more than one step

Framework	References	Detection	Evaluation	Influence
IRON	Morales *et al.* (2015)		System-based	Agent-based incentivization
EMIL	Andrighetto *et al.* (2010a, 2010b)	Agent-based observation	Agent-based salience and utility	–
Recommender	Savarimuthu *et al.* (2013b)	Agent-based observation	Agent-based salience and life-cycle stage	Recommendation

a norm; individuals may only consider short time-scales, whereas the true effect of norm adoption may only become apparent over the long-term; heterogeneous agents will have different personal preferences, goals and beliefs.

Specific approaches could involve group argumentation-based negotiation in order to resolve individual differences and come to a group consensus about the value of a norm, a voting system to measure the support for a new norm proposal, or a market-based system to allow individuals to allocate resources in order to promote the use of their preferred norm (or discourage the spread of an unwanted one). There may be situations where a norm may be individually very useful if followed by a subset of the population, but less useful if globally adopted. In such a case the early adopters of the norm may need to resolve a tension between promoting it for the benefit of society and restricting it for personal gain. As a real world example, consider the existence of patents that reward creative entrepreneurship while restricting the global adoption of the results of that creativity. Such situations are more complex that those that allow open discussion, since allowing the wider society to gain knowledge of the beneficial norm may negate some of its value to the individual. The possibility of such a situation means that group valuations must cope with deception and issues of trust. Such research could draw upon work in political science and social policy for inspiration, as these are challenges faced by human organizations and societies.

Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF under Award No. FA9550-15-1-0092.

References

Andrighetto, G., Campenni, M., Cecconi, F. & Conte, R. 2010a. The complex loop of norm emergence: a simulation model. In *Simulating Interacting Agents and Social Phenomena, Agent-Based Social Systems* **7**, Takadama K., Cioffi-Revilla C., Deffuant G. (eds). Springer, 19–35.

Andrighetto, G., Villatoro, D. & Conte, R. 2010b. Norm internalization in artificial societies. *AI Communications* **23**(4), 325–339.

Artikis, A. 2009. Formalising dynamic protocols for open agent systems. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Law*, ICAIL '09, 68–77. ACM.

Artikis, A. 2012. Dynamic specification of open agent systems. *Journal of Logic and Computation* **22**(6), 1301–1334.

Atzori, L., Iera, A. & Morabito, G. 2010. The internet of things: a survey. *Computer Networks* **54**(15), 2787–2805.

Axelrod, R. 1986. An evolutionary approach to norms. *The American Political Science Review* **80**(4), 1095–1111.

Balke, T., De Vos, M. & Padget, J. 2012. Normative run-time reasoning for institutionally-situated BDI agents. In *Coordination, Organizations, Institutions, and Norms in Agent System VII*, Cranefield S., van Riemsdijk M.B., Vázquez-Salceda J., Noriega P. (eds). Springer, 129–148.

Bedau, M. A. 1997. Weak emergence. *Noûs* **31**(s11), 375–399.

Berger, T., Schreinemachers, P. & Woelcke, J. 2006. Multi-agent simulation for the targeting of development policies in less-favored areas. *Agricultural Systems* **88**(1), 28–43.

Bettenhausen, K. & Murnighan, J. K. 1985. The emergence of norms in competitive decision-making groups. *Administrative Science Quarterly* **25**(4), 350–372.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bou, E., López-Sánchez, M. & Rodrguez-Aguilar, J. A. 2007. Towards self-configuration in autonomic electronic institutions. In *Coordination, Organizations, Institutions, and Norms in Agent Systems II*, Lecture Notes in Computer Science **4386**, 229–244. Springer.

Broughton, J. 1990. *Restraint Use by Car Occupants*. Transport and Road Research Laboratory Research Report, 289.

Bryant, V. 1985. *Metric Spaces*. Cambridge University Press.

Bunn, D. W. & Oliveira, F. S. 2001. Agent-based simulation: an application to the new electricity trading arrangements of England and Wales. *IEEE Transactions on Evolutionary Computation* **5**(5), 493–503.

Campillo-Sanchez, P. & Gomez-Sanz, J. J. 2015. A framework for developing multi-agent systems in ambient intelligence scenarios. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 1949–1950. International Foundation for Autonomous Agents and Multiagent Systems.

Casari, M. & Luini, L. 2009. Cooperation under alternative punishment institutions: an experiment. *Journal of Economic Behavior & Organization* **71**(2), 273–282.

Castelfranchi, C., Conte, R. & Paolucci, M. 1998. Normative reputation and the costs of compliance. *Journal of Artificial Societies and Social Simulation* **1**(3), 3.

Castelfranchi, C. & Falcone, R. 1998. Principles of trust for MAS: cognitive anatomy, social importance, and quantification. In *The Proceedings of the International Conference on Multi Agent Systems*, 72–79. IEEE.

Castelfranchi, C. & Falcone, R. 2003. Founding autonomy: the dialectics between (social) environment and agent's architecture and powers. In *Agents and Computational Autonomy*, Lecture Notes in Computer Science **2969**, 40–54. Springer.

Chen, C.-C., Nagl, S. B. & Clack, C. D. 2007. Specifying, detecting and analysing emergent behaviours in multi-level agent-based simulations. In *Proceedings of the 2007 Summer Computer Simulation Conference*, 969–976. Society for Computer Simulation International.

Chen, C.-C., Nagl, S. B. & Clack, C. D. 2009. A formalism for multi-level emergent behaviours in designed component-based systems and agent-based simulations. In *From System Complexity to Emergent Properties*, Aziz-Alaoui M.A., Bertelle C. (eds). Springer, 101–114.

Conte, R. & Castelfranchi, C. 1999. From conventions to prescriptions. Towards an integrated view of norms. *Artificial Intelligence and Law* **7**(4), 323–340.

Cranefield, S., Savarimuthu, B., Meneguzzi, F. & Oren, N. 2015. A Bayesian approach to norm identification. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 1743–1744. International Foundation for Autonomous Agents and Multiagent Systems.

Criado, N., Argente, E. & Botti, V. 2011. Open issues for normative multi-agent systems. *AI Communications* **24**(3), 233–264.

Criado, N., Argente, E., Noriega, P. & Botti, V. 2010. Towards a normative BDI architecture for norm compliance. In *Proceedings of the Multi-Agent Logics, Languages, and Organisations Federated Workshops*, **2010**, 65–81.

Deguet, J., Demazeau, Y. & Magnin, L. 2006. Elements about the emergence issue: a survey of emergence definitions. *Complexus* **3**(1–3), 24–31.

Delgado, J. 2002. Emergence of social conventions in complex networks. *Artificial Intelligence* **141**(1), 171–185.

Delgado, J., Pujol, J. M. & Sangüesa, R. 2003. Emergence of coordination in scale-free networks. *Web Intelligence and Agent Systems* **1**, 131–138.

De Pelsmacker, P. & Janssens, W. 2007. The effect of norms, attitudes and habits on speeding behavior: scale development and model building and estimation. *Accident Analysis & Prevention* **39**(1), 6–15.

Dignum, F., Dignum, V. & Jonker, C. M. 2009. Towards agents for policy making. In *Multi-Agent-Based Simulation IX*, Lecture Notes in Computer Science **5269**, 141–153. Springer.

dos Santos Neto, B. F., da Silva, V. T. & de Lucena, C. J. P. 2012. An architectural model for autonomous normative agents. In *Advances in Artificial Intelligence-SBIA 2012*, Barros L. N., Finger M., Pozo A. T., Gimenénez-Lugo G. A, Castilho M. (eds). Springer, 152–161.

Elhag, A. A., Breuker, J. A. & Brouwer, P. W. 2000. On the formal analysis of normative conflicts. *Information & Communications Technology Law* **9**(3), 207–217.

Emons, W. 2007. Escalating penalties for repeat offenders. *International Review of Law and Economics* **27**(2), 170–178.

Faillo, M., Grieco, D. & Zarri, L. 2013. Legitimate punishment, feedback, and the enforcement of cooperation. *Games and Economic Behavior* **77**(1), 271–283.

Finnemore, M. & Sikkink, K. 1998. International norm dynamics and political change. *International Organization* **52** (04), 887–917.

Fitzek, F. H. P. & Katz, M. D. 2007. Cellular controlled peer to peer communications: overview and potentials. In *Cognitive Wireless Networks*, Fitzek F.H.P., Katz M.D. (eds). Springer, 31–59.

Franks, H., Griffiths, N. & Jhumka, A. 2013. Manipulating convention emergence using influencer agents. *Autonomous Agents and Multi-Agent Systems* **26**(3), 315–353.

Galán, J. M. & Izquierdo, L. R. 2005. Appearances can be deceiving: lessons learned re-implementing Axelrod's 'evolutionary approach to norms'. *Journal of Artificial Societies and Social Simulation* **8**(3), 2.

Garlick, M. & Chli, M. 2009. The effect of social influence and curfews on civil violence. In *Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems – Volume 2*, 1335–1336. International Foundation for Autonomous Agents and Multiagent Systems.

Gibbs, J. P. 1965. Norms: the problem of definition and classification. *American Journal of Sociology* **70**(5), 586–594.

Gneezy, U., Meier, S. & Rey-Biel, P. 2011. When and why incentives (don't) work to modify behavior. *The Journal of Economic Perspectives* **25**(4), 191–209.

Griffiths, N. & Luck, M. 2010. Changing neighbours: improving tag-based cooperation. In *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems: Volume 1*, 249–256. International Foundation for Autonomous Agents and Multiagent Systems.

Günay, A. & Yolum, P. 2013. Constraint satisfaction as a tool for modeling and checking feasibility of multiagent commitments. *Applied Intelligence* **39**(3), 489–509.

Hart, H. L. A. 2012. *The Concept of Law*. Oxford University Press.

Haynes, C., Miles, S. & Luck, M. 2014. Monitoring the impact of norms upon organisational performance: a simulation approach. In *Coordination, Organizations, Institutions, and Norms in Agent Systems IX*, Lecture Notes in Computer Science **8386**, 103–119. Springer.

Hoffmann, M. J. 2005. Self-organized criticality and norm avalanches. In *Proceedings of the Symposium on Normative Multi-Agent Systems, AISB05: Social Intelligence and Interaction in Animals, Robots and Agents*, 117–125.

Holland, J. H. 1992. Complex adaptive systems. *Daedalus* **121**(1), 17–30.

Hollander, C. D. & Wu, A. S. 2011. The current state of normative agent-based systems. *Journal of Artificial Societies and Social Simulation* **14**(2), 6.

Itaiwi, A. M. K., Ahmad, M. S., Tang, A. Y. C. & Mahmoud, M. A. 2014. A proposed norms' benefits awareness framework for norms adoption. In *International Conference on Information Technology and Multimedia*, pages 287–292.

Joseph, S., Sierra, C. & Schorlemmer, M. 2008. A coherence based framework for institutional agents. In *Coordination, Organizations, Institutions, and Norms in Agent Systems III*, Lecture Notes in Computer Science **4870**, 287–300. Springer.

Kittock, J. E. 1994. Emergent conventions and the structure of multi-agent systems. In *Lectures in Complex Systems: The Proceedings of the 1993 Complex Systems Summer School, Santa Fe Institute Studies in the Sciences of Complexity Lecture Volume VI, Santa Fe Institute*, Nadel L. & Stein D. (eds). Addison-Wesley Publishing Company, 1–14.

Kotsiantis, S. & Kanellopoulos, D. 2006. Association rules mining: a recent overview. *GESTS International Transactions on Computer Science and Engineering* **32**(1), 71–82.

Kubík, A. 2003. Toward a formalization of emergence. *Artificial Life* **9**(1), 41–65.

Mahmoud, M. A., Ahmad, M. S., Ahmad, A., Yusoff, M. Z. M. & Mustapha, A. 2012a. A norms mining approach to norms detection in multi-agent systems. In *International Conference on Computer Information Science, Volume 1*, 458–463.

Mahmoud, S., Barakat, L., Miles, S., Taweel, A., Delaney, B. & Luck, M. 2014. Information-based incentivisation when rewards are inadequate. In *ECAI*, 591–596.

Mahmoud, S., Griffiths, N., Keppens, J. & Luck, M. 2010. An analysis of norm emergence in Axelrod's model. In *Proceedings of 8th European Workshop on Multi-Agent Systems*.

Mahmoud, S., Griffiths, N., Keppens, J. & Luck, M. 2012b. Efficient norm emergence through experiential dynamic punishment. In *Proceedings of the Twentieth European Conference on Artificial Intelligence, Volume 12*, 576–581.

Mahmoud, S., Griffiths, N., Keppens, J. & Luck, M. 2013. Norm emergence through dynamic policy adaptation in scale free networks. In *Coordination, Organizations, Institutions, and Norms in Agent Systems VIII*, Aldewereld H. & Sichman J. (eds), Lecture Notes in Computer Science 7756, 123–140. Springer Berlin Heidelberg.

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

Mahmoud, S., Miles, S., Taweel, A., Delaney, B. & Luck, M. 2015b. Norm establishment constrained by limited resources. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 1819–1820. International Foundation for Autonomous Agents and Multiagent Systems.

Miceli, T. J. & Bucci, C. 2005. A simple theory of increasing penalties for repeat offenders. *Review of Law & Economics* **1**(1), 71–80.

Mill, J. S. 1863. *Utilitarianism*. Parker, Son and Bourn.

Mintzberg, H. & Waters, J. A. 1985. Of strategies, deliberate and emergent. *Strategic Management Journal* **6**(3), 257–272.

Modgil, S., Oren, N., Faci, N., Meneguzzi, F., Miles, S. & Luck, M. 2015. Monitoring compliance with e-contracts and norms. *Artificial Intelligence and Law* **23**(2), 161–196.

Mogul, J. C. 2006. Emergent (mis) behavior vs. complex software systems. *ACM SIGOPS Operating Systems Review* **40**(4), 293–304.

Morales, J., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Vasconcelos, W. & Wooldridge, M. 2015. Online automated synthesis of compact normative systems. *ACM Transactions on Autonomous and Adaptive Systems* **10**(1), 2.

Mungovan, D., Howley, E. & Duggan, J. 2011. The influence of random interactions and decision heuristics on norm evolution in social networks. *Computational and Mathematical Organization Theory* **17**(2), 152–178.

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

Parunak, H. V. D. & VanderBok, R. S. 1997. Managing emergent behavior in distributed control systems. In *Proceedings of ISA Tech'97, Instrument Society of America*, 1–8.

Riveret, R., Artikis, A., Busquets, D. & Pitt, J. 2014. Self-governance by transfiguration: from learning to prescriptions. In *International Conference on Deontic Logic in Computer Science*, 177–191. Springer.

Savarimuthu, B. & Cranefield, S. 2011. Norm creation, spreading and emergence: a survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems–An International Journal* **7**(1), 21–54.

Savarimuthu, B., Cranefield, S., Purvis, M. & Purvis, M. 2008. Role model based mechanism for norm emergence in artificial agent societies. *Coordination, Organizations, Institutions, and Norms in Agent Systems III* **487**, 203–217.

Savarimuthu, B., Cranefield, S., Purvis, M. & Purvis, M. 2010. A data mining approach to identify obligation norms in agent societies. In *Agents and Data Mining Interaction*, Lecture Notes in Computer Science **5980**, 43–58. Springer.

Savarimuthu, B., Cranefield, S., Purvis, M. & Purvis, M. 2013a. Identifying prohibition norms in agent societies. *Artificial Intelligence and Law* **21**(1), 1–46.

Savarimuthu, B. & Dam, H. K. 2013. Towards mining norms in open source software repositories. In *International Workshop on Agents and Data Mining Interaction*, 26–39. Springer.

Savarimuthu, B., Padget, J. & Purvis, M. 2013b. Social norm recommendation for virtual agent societies. In *International Conference on Principles and Practice of Multi-Agent Systems*, 308–323. Springer.

Savarimuthu, B. T. R., Cranefield, S., Purvis, M. & Purvis, M. 2007. Norm emergence in agent societies formed by dynamically changing networks. In *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, 464–470.

Sen, O. & Sen, S. 2010. Effects of social network topology and options on norm emergence. In *Coordination, Organizations, Institutions and Norms in Agent Systems V*, Lecture Notes in Computer Science **6069**, 211–222. Springer.

Sen, S. & Airiau, S. 2007. Emergence of norms through social learning. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, 1507–1512.

Sensoy, M., Norman, T. J., Vasconcelos, W. W. & Sycara, K. 2012. OWL-POLAR: a framework for semantic policy representation and reasoning. *Web Semantics: Science, Services and Agents on the World Wide Web* **12**, 148–160.

Seth, A. K. 2008. Measuring emergence via nonlinear Granger causality. *Artificial Life* **11**, 545–552.

Shoham, Y. & Tennenholtz, M. 1995. On social laws for artificial agent societies: off-line design. *Artificial Intelligence* **73**(1–2), 231–252.

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

Singh, M. P. 2013. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* **5**(1), 21.

Singh, M. P., Arrott, M., Balke, T., Chopra, A. K., Christiaanse, R., Cranefield, S., Dignum, F., Eynard, D., Farcas, E., Fornara, N. & Gandon, F. 2013. The uses of norms. In *Normative Multi-Agent Systems, Volume 4*, Andrighetto G., Governatori G., Noriega P. & van der Torre L. (eds). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 191–229.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A. & Katz, Y. 2007. Pellet: a practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web* **5**(2), 51–53.

Sycara, K. P. 1998. Multiagent systems. *AI magazine* **19**(2), 79–92.

Tinnemeier, N., Dastani, M. & Meyer, J.-J. 2010. Programming norm change. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1*, 957–964. International Foundation for Autonomous Agents and Multiagent Systems.

Van der Aalst, W. M. P., van Dongen, B. F., Herbst, J., Maruster, L., Schimm, G. & Weijters, A. J. M. M. 2003. Workflow mining: a survey of issues and approaches. *Data & Knowledge Engineering* **47**(2), 237–267.

Vasconcelos, W., Kollingbaum, M. J. & Norman, T. J. 2007. Resolving conflict and inconsistency in norm-regulated virtual organizations. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems*, 632–639. ACM.

Vasconcelos, W. W., Kollingbaum, M. J. & Norman, T. J. 2009. Normative conflict resolution in multi-agent systems. *Autonomous Agents and Multi-Agent Systems* **19**(2), 124–152.

Villatoro, D., Andrighetto, G., Sabater-Mir, J. & Conte, R. 2011a. Dynamic sanctioning for robust and cost-efficient norm compliance. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Volume 11*, 414–419.

Villatoro, D., Sabater-Mir, J. & Sen, S. 2011b. Social instruments for robust convention emergence. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Volume 11*, 420–425.

Villatoro, D., Sen, S. & Sabater-Mir, J. 2009. Topology and memory effect on convention emergence. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 233–240. IEEE Computer Society.

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

Wolf, T. D. & Holvoet, T. 2005. Emergence versus self-organisation: different concepts but promising when combined. In *Engineering Self-Organising Systems*, Brueckner S., Serugendo G. D. M., Karageorgos A. & Nagpal R. (eds), Lecture Notes in Computer Science 3464, 77–91. Springer.

y López, F. L. & Luck, M. 2004. A model of normative multi-agent systems and dynamic relationships. In *Regulated Agent-Based Social Systems*, Lindemann G., Moldt D. & Paolucci M. (eds), Lecture Notes in Computer Science **2934**, 259–280. Springer.

Zhang, Y. & Leezer, J. 2009. Emergence of social norms in complex networks. In *Proceedings of the International Conference on Computational Science and Engineering, Volume 4*, 549–555. IEEE.

Zimmermann, M. G. & Eguluz, V. M. 2005. Cooperation, social networks, and the emergence of leadership in a prisoner's dilemma with adaptive local interactions. *Physical Review E* **72**(5), 056118.