MODELING NORM DYNAMICS IN MULTI-AGENT SYSTEMS

CHRISTOPHER K. FRANTZ
Norwegian University of Science and Technology (NTNU)
Department of Computer Science (IDI)
2815 Gjøvik, Norway
cf@christopherfrantz.org

GABRIELLA PIGOZZI
Université Paris-Dauphine, PSL Research University
CNRS, LAMSADE
75016 Paris, France
gabriella.pigozzi@dauphine.fr

1 Introduction and Motivation

Since multi-agent systems are inspired by human societies, they do not only borrow their coordination mechanisms such as conventions and norms, but also need to consider the processes that describe *how norms come about*, *how they propagate in the society*, and *how they change over time*.

In the NorMAS community, this is best reflected in various norm life cycle conceptions that look at normative processes from a holistic perspective. While the earliest life cycle model emerged in the research field of international relations, the first life cycle model in the AI community has been proposed at the 2009 NorMAS Dagstuhl workshop by Savarimuthu and Cranefield [2009] and is based on a comprehensive survey of then existing contributions to the research field. Subsequently, two further models have been proposed that offer more refined accounts of the fundamental underlying processes.

In this article, we review all existing norm life cycle models (Section 2), including the introduction of the individual life cycle models and their contextualization with specific contributions that exemplify life cycle processes. In addition, we provide a *comprehensive*

We would like to thank Bastin Tony Roy Savarimuthu for bringing us together and providing us with the opportunity to collaborate on this article.

contemporary overview of individual contributions to the area of NorMAS and a *systematic comparison of the discussed life cycle models* (Section 2.6). Based on this analysis, we propose a refined *general norm life cycle model* that resolves terminological ambiguities and ontological inconsistencies of the existing models while reflecting the contemporary view on norm formation and emergence.

This comprehensive review of life cycle models represents the birds-eye perspective on dynamics in normative multi-agent systems, which is complemented by research areas that operate at the intersection of normative processes captured by life cycle models. In addition to this holistic perspective, we thus discuss two active research fields that deal with norm dynamics: norm change and norm synthesis.

In human societies, norms change over time: new norms can be created to face changes in the society, old norms can be retracted either because they became obsolete or because superseded by others, and also norms can be modified. Thus, multi-agent systems too need mechanisms to model and reason about norm change. The field of *norm change* (Section 3) puts a specific focus on the definition of mechanisms that describe and regulate the change of norms over time. Essential aspects include the *translation of legal to logical specifications*, the *definition of a normative approach to norm change*, and the *tuning of computational mechanisms for norm change*. This research area is rather recent and to date there is still no consensus on a common account for norm change. This section retraces the historical development and debates within this field and provides an outlook on future directions.

The second subfield, *norm synthesis* (Section 4), has a longer history that has its roots in the systems engineering domain and is concerned with the use of norms and social laws as scalable coordination mechanisms in open systems. The associated challenges are twofold and have led to the development of distinct branches, with one concentrating on the analysis of factors that mitigate the *emergence of norms or conventions*, and the second one focusing on the *identification and classification of norms* in existing normative environments. This section identifies a taxonomy of norm synthesis approaches based on a comprehensive literature overview of the field, and illustrates contemporary developments using selected contributions.

We conclude this article by contextualizing the discussed subfields with the proposed general norm life cycle model, reflecting on the progression of research on norm dynamics, and finally, by providing an outlook on contemporary and future challenges of modeling of norm dynamics.

2 Norm Life Cycle Models

In the following sections, we introduce four norm life cycle models discussed in the literature to date. The models are organized chronologically, and, with exception of the last model by Mahmoud *et al.* (Section 2.4), are of increasing complexity. The first model by Finnemore and Sikkink (Section 2.1) describes normative processes to capture the dynamics of international relations, whereas the models by Savarimuthu and Cranefield (Section 2.2), Hollander and Wu (Section 2.3), and Mahmoud *et al.* (Section 2.4) have been proposed in the research field of normative multi-agent systems. Since the identified individual processes that constitute all models are supported by relevant literature contributions, we provide an updated review of associated literature. The later three models represent incremental extensions of earlier models, and, in consequence, feature redundant elementary processes. In such cases, we refer the reader to the corresponding processes in earlier life cycle models.

2.1 Model 1: Finnemore & Sikkink

2.1.1 Overview

Norms have been traditionally studied in the social sciences [Crawford and Ostrom, 1995] (see also Finnemore and Sikkink [1998], Elster [1989], Bicchieri [2006]), but no consensus yet exists on how norms emerge and are subsequently adopted in a society. In order to understand the role that norms play in international politics, Finnemore and Sikkink [1998] introduced the concept of “life cycle" to model the origin and the dynamics of norms. They claimed that norms follow a specific pattern and that each portion of the life cycle is characterized by different actors and mechanisms. The term of life cycle was later imported and became particularly relevant for the study and modelling of normative multi-agent systems.

Finnemore and Sikkink's norm life cycle is a three-stage process, as shown in Figure 1: the first step is norm *emergence*, followed by norm *acceptance* (following Sunstein [1996], also called norm *cascade*), and the last stage is norm *internalization*. The move from norm emergence to norm cascade happens once the norm has been accepted by a certain amount of actors (the threshold point).

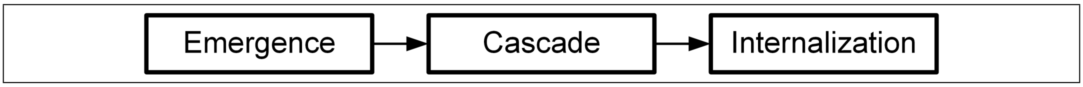

Figure 1: Finnemore and Sikkink's Norm Life Cycle Model

It is important to mention that a norm does not necessarily complete a life cycle. If, for instance, a norm does not reach the threshold point, it will not move from the emergence stage to the cascade stage. The different stages of Finnemore and Sikkink ’s model are supported by examples coming from women’s movement of suffragettes and laws of war.

2.1.2 Stage 1: Norm Emergence

At the origin of norms we find norm *entrepreneurs*, agents committed to persuade a critical mass to support new norms or to alter existing ones in order to achieve desirable behaviour in a state or community. As Hoffmann [2003] notes, leaders and entrepreneurs are not novel concepts in political science [Nadelman, 1990; Young, 1990; Schneider and Teske, 1992; Bianco and Bates, 1990]: “Entrepreneurship is a popular factor for explaining solutions to collective action problems, equilibrium choice, the emergence of cooperation as well as norms" (Hoffmann [2003], p. 8). As an example of a norm entrepreneur, Finnemore and Sikkink mention Henry Dunant, who played a crucial role in forming the norm that, in time of war, doctors and wounded soldiers should be treated as noncombatants and, by consequence, granted immunity.

The task of norm promoters is rarely easy. More often proposing a new norm implies competing with existing social contexts and established states of affairs. This means that one has to be ready to battle with competing norms or conflicting interests. The mechanisms by which individuals manage to convince other individuals is debated [Checkel, 1998; Risse and Sikkink, 1999]. Finnemore and Sikkink argue that the difficulty of the task explains why norm entrepreneurs frequently resumed to controversial or even illegal acts (such as the protests engaged by suffragettes, who refused to pay taxes and went on hunger strikes, among other things). Altruism, empathy and commitment to an ideal are the motives that Finnemore and Sikkink attribute to norm entrepreneurs to explain their dedication.

Observing norm emergence in international relations, Finnemore and Sikkink stress that norm entrepreneurs act within organizational platforms, like nongovernmental organisations. This facilitates the reaching of the threshold point and thus the emergence of the norm. In the context of international politics, empirical studies fix such threshold around one-third of the total states, even though some states are more critical to the adoption of a norm than others. The second stage (norm cascade) is reached when the threshold is passed.

Subsequent models, like Hollander and Wu [2011b], will refine Finnemore and Sikkink ’s norm life cycle and will replace entrepreneurs by machine learning and cognitive approaches (Section 2.3).

2.1.3 Stage 2: Norm Cascade

We have seen that once the threshold of the critical mass is passed, according to the Finnemore and Sikkink’s model, we move to the stage of norm cascade. This is called so because the acceptance rate of the new norm among the individuals increases rapidly. The mechanism that seems to govern the acceptance of a norm is *socialization*, a kind of persuasion by some agents to others to embrace a certain norm. In the case of states, such persuasion appears to lean against the need of a state to be recognized as a member of an international organisation. In other words, exactly as it happens to people, countries would be exposed to peer pressure. In particular, the desire to acquire or increment internal and international legitimation, the pressure of conformity and the need for norm leaders to increase their esteem seem to be the reason to respond to such a pressure.

2.1.4 Stage 3: Internalization

If a norm reaches the third and last stage, it becomes internalized. This means that such norm is acquired and not object of debate anymore. As Epstein stated, once a norm is accepted, people “conform without really thinking about it" (Epstein [2001], p.1). Examples of nowadays internalized norms are the abolition of slavery or the right to vote for women. But internalized norms can also be specific to certain professions. Finnemore and Sikkink mention the examples of doctors and soldiers, who become acquainted with different “normative biases": “Doctors are trained to value life above all else. Soldiers are trained to sacrifice life for certain strategic goals" (Finnemore and Sikkink [1998], p.905).

2.1.5 Discussion

Constructivists (to which Finnemore and Sikkink’s approach belongs) have been criticized for failing to account how entrepreneurs hammer new norms or come to propose the alteration of existing ones, as well as how they manage to convince other critical agents in their vision. Hoffmann [2003] partially addresses such criticisms by building an agent-based model to explore the role of norm entrepreneurs. His model does not tackle the question of how entrepreneurs convince other agents, but focuses “on the unexamined assumption that a persuasive entrepreneur can influence the outcomes that arise from the interactions of heterogeneous, interdependent agents" (Hoffmann [2003], p. 13). His model shows that the constructivist’s hypothesis of the role of norm entrepreneurs is indeed plausible. In particular, his aim is to understand under what conditions a norm entrepreneur can function as a norm catalyser for the emergence of new norms and the alteration of existing ones. Norm entrepreneurs turn out to be able to influence norm emergence even when they can reach only a small portion of the population (around 30%), and their influence increases with their reach. Hoffmann’s model suffers (as the author himself acknowledges) from some limitations, like the assumption of a unique norm entrepreneur, the lack of communication among agents, agents’ power is not modelled, and only non-complex norms are considered.

2.2 Model 2: Savarimuthu & Cranefield

2.2.1 Overview

The first life cycle model for norms we have encountered was proposed in the context of international relations. As we have seen, Finnemore and Sikkink [1998] directed their attention to human societies and to processes that can explain how norms emerge and spread within and among states. Ten years separate Finnemore and Sikkink's work from the second model we consider here, the life cycle model proposed by Savarimuthu and Cranefield [2009; 2011].

Savarimuthu and Cranefield's model comes from the study of simulation-based works on norms in the context of multi-agent systems. By looking at the various mechanisms employed by the researchers working on simulation on norms, they extend the three-stage model of Finnemore and Sikkink.

Savarimuthu and Cranefield's contribution came in two papers: the first one [Savarimuthu and Cranefield, 2009] presented a four phases norm life cycle (*norm creation, spreading, enforcement and emergence*), whereas the subsequent [Savarimuthu and Cranefield, 2011] included one additional stage (*identification*). For this reason, in the present section we will focus on the latter, more recent, contribution. For each step Savarimuthu and Cranefield provide a categorisation of the mechanisms that have been employed in the simulation-based works on norms, as shown in Figure 2.

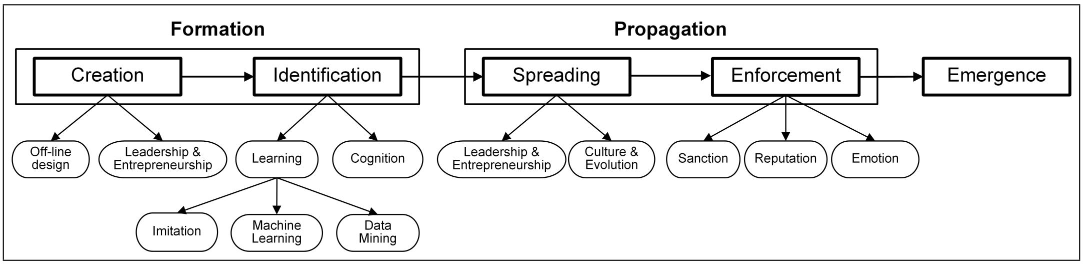

Figure 2: Savarimuthu and Cranefield's Norm Life Cycle Model

2.2.2 Norm Creation

Unlike Finnemore and Sikkink [1998] , who acknowledged only the role of norm entrepreneurs for the creation of norms, Savarimuthu and Cranefield [2011] realize that in the context of multi-agent systems, norms can be created by three different approaches: off-line design, norm leaders and norm entrepreneurs. In off-line design the norm is introduced by an external designer and is hard-wired into the agents. Norm leaders, on the other hand, are powerful agents of the system that (following a democratic or an authoritarian process) create norms for the other agents to follow. Finally, norm entrepreneurs are not necessarily norm leaders. Similarly, as seen in Finnemore and Sikkink's model, an entrepreneur can propose a new norm that he thinks is beneficial to the society. But until the entrepreneur does not succeed to persuade the other agents to accept such norm, the norm is not a social norm.

Off-line Design One of the most well-known works in the area of off-line design is Shoham and Tennenholtz [1995]'s work on synthesising social laws, specifically in the traffic domain. In this specific context, off-line design implies that mobile robots (as traffic participants) are initialized with a set of traffic laws ('rules of the road') that have been computed at design time in order to prevent collisions at runtime. Such rules allow to minimize the need of a central coordinator on the one hand and that of a negotiation mechanism among agents on the other hand. Traffic laws provide the agents with a set of social laws that help them avoiding collisions. A multi-robot grid system is considered, where m robots can move on an $n \times n$ grid. Shoham and Tennenholtz suggest one can imagine rows and columns of that grid as lanes in a supermarket. In order to avoid the collision between robots (which happens when more than one robot occupies the same coordinate), some traffic laws are given. For instance, one may impose that in odd rows agents can move only right, in even rows they can move only left, in odd columns they can move only down and in even columns the only movement possible is up. Rules define also the priority when two or more robots approach a junction and how robots can change their movement direction Shoham and Tennenholtz [1995]'s work has subsequently been extended (e.g. to consider the minimality of social laws [Fitoussi and Tennenholtz, 2000]) and has found various adaptations in works on norm emergence (e.g. [Sen and Airiau, 2007; Mukherjee *et al.*, 2007]).

A similarly influential model from the sociological domain is Conte and Castelfranchi [1995b]'s evaluation of norms for the purpose of aggression control to facilitate cooperation in a stylized food-gathering society. In their model societies are selectively initialized as either *strategic* or *normative*, where strategic agents systematically attack fellow food-carrying agents, whereas normative ones accept a notion of possession, thus promoting a higher level of survival at the macro level. Results have shown that normative populations do better than strategic ones. However, in mixed populations strategic agents do much better than the normatives. The reason is that non-normative agents benefit from the behaviour of normatives.

Castelfranchi *et al.* [1998] have further extended the model to consider the role of reputation (see Section 2.2.5). The role of reputation is considered also in Hales [2002], which extended Castelfranchi *et al.*'s food-consumption problem by assigning agents to the group of normative agents or to the group of cheaters.

Walker and Wooldridge [1995] observe that the simplicity of off-line design models comes at a price. To be truly beneficial, such approach requires that all characteristics of a system should be known a priori (which is not the case for open systems, for example). Another difficulty is that it is extremely costly and time-consuming to constantly reprogram agents, which is required in case agents' goals change, as it happens in complex systems. Moreover, Savarimuthu and Cranefield [2011] note that it is not realistic to assume that all agents follow a given norm.

Leadership and Entrepreneurship Mechanisms Leaders are agents who have the social power and abilities to persuade other agents to accept a norm. Leadership mechanisms have been employed for norm emergence and norm spreading (see Section 2.2.4). Verhagen [2001] considers agents with a certain degree of autonomy and a normative advisor (as in Boman [1999]'s approach) from whom they receive comments on an agent's decision to follow or not to follow a norm. Once an agent decides to follow a specific norm, it announces it to the whole society. The normative advisor as well as other agents can send feedback to that agent, who may assign a greater weight to the comments received from the leader.

In Savarimuthu *et al.* [2008a] a society can have several normative advisors (or role models) who give advice to agents who are their followers. Agents are connected to each other through one social network topology among fully connected networks, random networks and scale-free networks. The interesting twist is that an agent can be at the same time a role model for some agents and a follower of some other agent. Since several norm leaders can exist, different norms can emerge in the society.

Norm entrepreneurs were notably introduced in Finnemore and Sikkink's norm life cycle model, presented in Section 2.1. Hoffmann [2003] has experimented on the notion of norm entrepreneurs, as seen in the Discussion subsection of Section 2.1.

2.2.3 Norm Identification

The first norm life cycle model proposed by Savarimuthu and Cranefield [2009] consisted of four stages (norm creation, spreading, enforcement and emergence). The idea being that, as in [Finnemore and Sikkink, 1998], once a norm is created, it may spread in a society if certain conditions are satisfied. However, in [Savarimuthu and Cranefield, 2011], they added the *identification* step between norm creation and spreading. Such step is needed in all those situations in which a norm has not been explicitly created, for example when a norm results from the interaction process among agents. In those cases, agents have first to be able to identify the created norms. Simulation-based works on norms have explored two approaches for norm identification: agents can learn new norms by imitation, machine learning or data mining mechanisms; alternatively, agents can use their cognitive abilities to infer and recognize the norms of a system.

Learning Mechanisms – Imitation Among the simulation models that experimented on learning mechanisms based on imitation is that of Epstein [2001]. Using a driving setting in which agents can observe whether other agents (within a certain radius) drive on the right or on the left, Epstein showed that agents conform to the driving preferences of the majority of the observed agents. Imitation mechanisms can explain the identification and the spreading of a norm.

Yet, some authors, like López y López and Márquez [2004] as well as Campenni *et al.* [2009], cast some doubts on the claim that such mechanisms can explain the co-existence of different norms in a group of agents. Instead of seeing norms are hard-wired in the agents, Campenni *et al.* [2009] imagine the interaction between agents coming from different societies. Their goal is to investigate the role of cognition in norm recognition: How do agents tell that something is a norm? In their model, there are four scenarios, some actions that are context-specific and one action that is common to all scenarios. In one set of simulations, agents can change contexts, whereas in another set of simulations, at a certain moment, agents must stay in the context they have reached and can interact only with agents that are in the same context (imagine a situation in which a population is split into two groups and each group is constrained to not have contacts with the other group). The purpose of this second set of simulations is to show that frequency may be a sufficient (but not necessary) condition for agents to converge to the same action. Results show that new norms can emerge, eventually giving rise to the competition between two rival norms.

Learning Mechanisms – Machine Learning Shoham and Tennenholtz [1992a] employed co-learning, a reinforcement learning mechanism that makes an agent choose the strategy that revealed to be the most successful in the past. They showed that norm emergence decreases with the decrease of the frequency of the updates of an agent's strategy. The efficiency of norm emergence turned out to decrease also with the increase of an agent's memory flush.

Building on the scenario introduced in [Conte and Castelfranchi, 1995b], Walker and Wooldridge [1995] ran 16 experiments with different parameters for the size of the majority and the update function (the latter could depend on the majority rule, on the memory flush or on communication mechanisms). Results showed that the network topology and communication may play an important role and, hence, more simulations are needed to better understand mechanisms for norm emergence.

More recently, norm emergence has been investigated using social learning in a model in which agents repeatedly interact with other agents by Sen and Airiau [2007]. Experiments took into account different population sizes, various learning strategies, and number of available actions. The specific situation is that of learning of which side of the road to drive on but also the problem of who has the priority if two agents gain a junction at the same time. The outcomes confirm that such a mode of learning is a robust mechanism for the emergence of social norms.

Learning Mechanisms – Data Mining An approach to norm identification that uses association rule mining to identify obligation norms is Savarimuthu *et al.* [2010b]'s *Obligation Norm Inference* (ONI) algorithm. Such model enables agents to sense their environment, memorize experiences and observations as well as normative signals, which build the basis for the identification of personal norms (p-norms) and group norms (g-norms). The memorized event episodes are then mined for obligation norms using association rules algorithms. The agent-based simulation experiment considers a virtual restaurant in which agents may not know whether the restaurant expects the customers to order and pay for the food at the counter before eating or if they are expected to order, consume the food and pay only before leaving. Another protocol agents may need to identify is the tipping norm: in some countries, for example, tipping is expected (in the USA, for instance), whereas in others (like most countries in Europe) it is not expected. The difficulty in identifying an obligation norm is that a sanction is triggered by the absence of an action (a customer in a restaurant may be sanctioned if he is not tipping the waiter). Savarimuthu *et al.* [2013a] propose a corresponding approach for the identification of prohibition norms.

Savarimuthu and Cranefield [2011] observe that data mining is a promising approach. However, explicit signals for sanctions or reward have to be present in order for norms to be easily identified.

Cognition The EMIL-A architecture [Andrighetto *et al.*, 2007; Campenni *et al.*, 2009; Andrighetto *et al.*, 2010][1] is a cognitive architecture to explore how agents' mental abilities may explain the acquisition of new norms. Reinforced candidate norms are identified from observed normative information (represented as normative frame) that traverses different memory layers, representing the transition from short-term experiences to long-term memory. Once established, normative beliefs are held in a *Normative Board*, along with associated normative action plans. These internalized normative beliefs inform the agent's goal generation, decision-making and action planning. The previously discussed work by Savarimuthu *et al.* [2010b] also proposed an architecture for agents to identify norms using agents' cognition abilities.

[1] Campenni *et al.* [2009]'s contribution is a notable extension of Andrighetto *et al.* [2010]'s work.

2.2.4 Norm Spreading

Once a norm has been explicitly created or agents have identified it, the norm can start being spread in the society. Among the different mechanisms that can serve this purpose, there are leadership and entrepreneurship that we already encountered in the norm creation stage, but also cultural and evolutionary mechanisms.

Culture and Evolution Cultural and evolutionary mechanisms have been considered in [Boyd and Richerson, 1985; Chalub *et al.*, 2006]. According to Boyd and Richerson [1985] social norms can be propagated along three types of transmissions: vertical, horizontal and oblique. *Vertical relationships* describe the intergenerational transmission of norms by parents to offspring, whereas *horizontal transmission* occurs among peers of a given generation. *Oblique relationships* combine the former two and describe the unidirectional dissemination of norms by authority figures towards their contemporary subalterns. Vertical relationships are constrained to the intergenerational sharing of norms which makes them particularly applicable to evolutionary models such as Axelrod's norm game [Axelrod, 1986]. Horizontal approaches assume a uniform social structure, which limits this approach to abstract group or society representations, as is the case for large parts of the norm emergence work (e.g. [Sen and Airiau, 2007; Villatoro *et al.*, 2011a; Mihaylov *et al.*, 2014; Airiau *et al.*, 2014]; Section 4). The last relationship type lends itself well to model inter- and intra-generational norm transmission for comprehensive society representations that consider power and authority structures. Examples for this include Franks *et al.* [2014]'s use of *Influencer Agents* to drive the norm convergence, or Yu *et al.* [2015]'s hierarchical approach to information sharing.

Savarimuthu and Cranefield [2011] note that if cultural and evolutionary mechanisms can explain how a norm is spread, they cannot answer the question of how a norm is internalized in the first place.

2.2.5 Norm Enforcement

The existence of a norm presupposes that such norm can be violated. Norm enforcement mechanisms serve to deter agents from violating a norm. This can be done through punishment, via some mechanisms that negatively affect the reputation of a norm violator, or again by affecting the agent's emotions (for example, by instilling a sense of guilt in the norm violator). Savarimuthu and Cranefield [2011] stress that norm enforcement can also play a role in the spreading process of a norm. Observing the punishment of a norm violator can either discourage other agents from violating that norm or identifying that norm, in case it was not explicitly created.

Sanctions The most well-known work on external sanctions is Axelrod [1986]'s norm game that specifically explores the notion of metanorms, i.e. the sanctioning of non-sanctioning observers of violations, to sustain a society's norm.[2] An essential challenge of normative regulation (in artificial systems as in real life) is the balance of cost and effect of sanctions, both to minimize the cost of enforcement, while maximizing the effect in order to regulate behaviour effectively [Axelrod, 1986; Horne, 2001; Savarimuthu *et al.*, 2008a]. Mahmoud *et al.* [2012; 2015] refine Axelrod's model by investigating the effect of dynamic punishment, and ultimately propose an alternative to Axelrod's evolutionary approach based on individual learning to produce a model in which norms can stabilize within a given generation.

In López y López [2002; 2003] a model where agents have goals and different personalities is developed. Punishments and rewards are considered only when they affect an agent's goals.

Reputation A positive or negative opinion about one agent from the interacting agents in a society can play a substantial role in the norm compliance in a group of agents.

In Castelfranchi *et al.* [1998]'s and Younger [2004]'s models, ostracism is an implicit result of reputation sharing, which leads to the exclusion of individuals from future interaction. In particular, Castelfranchi *et al.* [1998]'s game reconsiders Conte and Castelfranchi [1995b]'s stylized food-gathering society seen in Section 2.2.2, with the addition of normative reputation: agents learn the reputation of other agents, that is, they learn whether an agent is normative or strategic (i.e. a cheater). However, in order to be profitable, the information about cheaters must be communicated to other agents. In the context of multi-agent systems Perreau de Pinninck *et al.* [2010] propose a distributed mechanism that affords the isolation of violating nodes in the context of peer-to-peer applications. They evaluate its properties for various network topologies.

Emotion Staller and Petta [2001] introduce an extension of the cognitive agent architecture JAM [Huber, 1999] with components to augment the rational agent model with emotion appraisal processes, an aspect considered essential to mediate any form of norm enforcement [Scheve *et al.*, 2006]. Fix *et al.* [2006] propose a model of normative agents that include the display of emotional responses to normative actions. In this work the agents' internal states are represented using reference nets [Valk, 1998], a variant of Petri nets.

[2] Axelrod's contribution was impressive and extremely influential. However, it should be noted that Galan and Izquierdo [2005] have shown that his results are not stable. When running many more simulations of Axelrod's model and for longer, opposite results can be obtained. As the authors also stress, one should not forget that their analysis required computational power which was not available when Axelrod proposed his model.

2.2.6 Norm Emergence

Once a norm has spread across a certain proportion of the society (according to different simulation results, the minimum required is a third of the population), it is said that the norm has emerged. This implies that a significant proportion of the population recognizes and follows that norm. It is worth noticing, however, that such process can be reverted. A norm may lose its appeal in a group and is hence either abandoned, replaced or modified by a competing one.

No specific category of empirical work on norms is associated with norm emergence. However, there is one category whose impact is notable across all stages of norm development. This is the consideration of network topology, as described in the Transmission part in Section 2.3.2.

2.2.7 Discussion

Savarimuthu and Cranefield [2011]'s life cycle model is an extension of the life cycle introduced by Finnemore and Sikkink in [Finnemore and Sikkink, 1998]. There are, however, two main differences.

The first one is that, whereas Finnemore and Sikkink's model was thought for human societies, Savarimuthu and Cranefield direct their attention to normative multi-agent systems and to simulation studies of norms using software agents. The second difference is that Savarimuthu and Cranefield not only capture two additional steps in their model, but also that for each phase, they consider more mechanisms.

2.3 Model 3: Hollander & Wu

2.3.1 Overview

To date, the most complex norm life cycle model has been proposed by Hollander and Wu [2011b]. Their model refines the ones initially introduced by Finnemore and Sikkink [1998] (Section 2.1) and Savarimuthu and Cranefield [2011] (Section 2.2), resulting in a total of ten *normative processes*, namely *creation*, *transmission*, *recognition*, *enforcement*, *acceptance*, *modification*, *internalization*, *emergence*, *forgetting*, and *evolution*. In contrast to the earlier models, Hollander and Wu identify three superprocesses (*enforcement*, *internalization*, and *emergence*) that combine elementary processes and characterize their high-level function. Note that the superprocess labels are borrowed from the most essential elementary process out of all processes they combine. A further novelty is the interpretation of *emergence* as an iterative process, and *evolution* as a metaprocess the authors refer to as "end-to-end process" [Hollander and Wu, 2011b]. The schema in Figure 3 provides a systematic overview of the complete life cycle. Where existing, the superprocesses are represented as boxes comprising their elementary processes, with the corresponding superprocess label highlighted in bold font. We will briefly outline the entire life cycle before introducing the individual processes in greater detail.

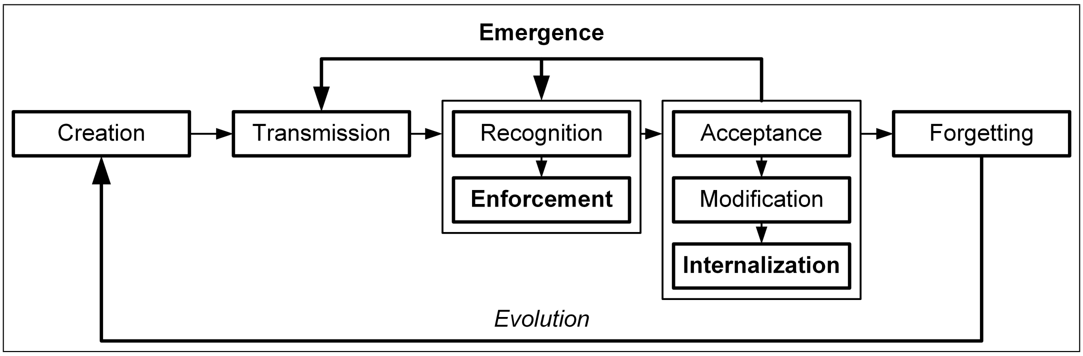

Figure 3: Hollander and Wu's Norm Life Cycle Model

Initially, potential norms are explicitly created, before being transmitted to the wider society, and rely on recognition and enforcement processes (captured in the superprocess *enforcement*) to promote their adoption. The superprocess *internalization* involves the decision whether to accept a norm, potentially modifying it, and finally, internalizing it, and thus becoming an enforcer of the norm itself. The subsequent cyclic reinforcement of the norm, including transmission, enforcement and internalization (tagged *emergence*), determines whether the initial *potential norm* becomes a norm. If attaining normative status, norms undergo a continuous refinement that requires reiteration through the elementary processes to gain salience. Any norm modification, such as the adaptation to new circumstances, implies that some normative content is forgotten. Swipe-card payments for bus services, for example, make it increasingly permissible for individuals to enter buses through arbitrary doors, instead of requiring the traditional entry through specific doors for payment. Contrasting the gradual forgetting of normative content, norms can be superseded by alternative norms, in which case the original norm is forgotten in its entirety. For example, over the past decades in many Western countries the general tolerance towards smoking in public places has been progressively replaced with general rejection.

In the following, we will discuss selected processes in greater detail and contextualize those with the earlier life cycle models as well as recent developments.

2.3.2 Life Cycle Processes

Creation Similar to Savarimuthu and Cranefield [2011], Hollander and Wu acknowledge that norm creation involves a wide range of different processes, including methods found in the natural world [Boella *et al.*, 2008; Finnemore and Sikkink, 1998; López y López *et al.*, 2007; Savarimuthu and Cranefield, 2009], such as spontaneous emergence from social interaction, decree by an agent in power, or negotiation within a group of agents. However, in the context of work on NorMAS, Hollander and Wu identify two primary methods of norm creation, namely off-line design [Conte and Castelfranchi, 1995a; Shoham and Tennenholtz, 1995] and autonomous innovation [Hollander and Wu, 2011b]. While off-line design assumes that experimenters create the norms a priori and inject those into instantiated agents, autonomous innovation (akin to 'on-line design') assigns the role of norm creation to agents themselves.

Notable works in the area of off-line design include Shoham and Tennenholtz [1995] and Conte and Castelfranchi [1995b], as already discussed in Section 2.2.2.

Autonomous innovation covers a broader range of approaches, ranging from the adoption of specific strategies to the challenging problem of ideation, namely giving agents the ability to produce novel ideas without external input.

In contrast to previous models' norm creation mechanisms in the form of norm leadership/entrepreneurship (see Sections 2.1 and 2.2), Hollander and Wu [2011b] identify two types of mechanisms used for autonomous innovation, namely:

- Game-theoretical and machine learning approaches (e.g. Sen and Airiau [2007], Mukherjee *et al.* [2007], Perreau de Pinninck *et al.* [2008], Urbano *et al.* [2009], Sen and Sen [2010], Savarimuthu *et al.* [2010b]), and
- Cognitive approaches (e.g. Savarimuthu *et al.* [2010b], Andrighetto *et al.* [2007]).

Even though many models use a combination of those mechanisms,[3] their application tends to serve distinctive purposes. Game-theoretical approaches emphasize the identification of optimal strategies from a set of given strategies, thus representing an incremental step from off-line design towards autonomous norm innovation. Machine learning is generally used in conjunction with game-theoretical approaches, mostly to represent a notion of memory (e.g. Sen and Airiau [2007], Mukherjee *et al.* [2007]).

Essential work that combines game-theoretical and machine learning approaches is the research field of *norm emergence* or *convention emergence*. This field concentrates on the identification of factors that promote high convergence levels for norms within the observed society. While decision-making itself is modelled as some form of game (with 'rules of the road' [Shoham and Tennenholtz, 1995] as the preferred coordination game), agent components such as memory are represented using machine learning (commonly reinforcement

[3] Examples for combining game-theoretical and machine learning approaches are provided by Sen and Airiau [2007] and Mukherjee *et al.* [2007]; an example for the combined use of machine learning and cognitive approaches is Savarimuthu *et al.* [2010b]'s work.

learning in the form of Q-learning [Watkins and Dayan, 1992]). Depending on the aspect of interest, the model is augmented with additional mechanisms to investigate the influence of memory (e.g. Villatoro *et al.* [2009]), characteristics of network topologies and structural dynamics (e.g. Savarimuthu *et al.* [2007], Villatoro *et al.* [2009], Sen and Sen [2010], Villatoro *et al.* [2013]), norm transmission mediated by social learning (e.g. Sen and Airiau [2007], Mukherjee *et al.* [2007; 2008], Airiau *et al.* [2014]), as well as adaptive sanctioning (e.g. Mahmoud *et al.* [2012; 2015]).

Sen and Airiau [2007], for example, let agents engage in social interaction in the context of the 'rules of the road' scenario (described in Section 2.2), in which cars approach an unregulated intersection and have to identify an optimal coordination mechanism, such as 'yield to the right', and prevent deadlocks (both cars yield) or collision.[4] Agents memorize past encounters and adjust their behaviour based on the success of their action. As part of their evaluation, Sen and Airiau explore different population sizes, action spaces and learning algorithms to show how agent societies can autonomously arrive at stable norms.

Further approaches investigate the influence of hierarchical structures on the distribution of norms (e.g. Franks *et al.* [2013; 2014], Yu *et al.* [2013; 2015]).[5]

While work in the area of norm emergence concentrates on the interactions and corresponding macro-level outcomes, cognitive approaches concentrate on the mechanics of normative agent architectures. Cognitive norm architectures contextualize perceived behaviour with existing beliefs to infer normative content and/or consider normative beliefs in their deliberation process. Approaches of this kind generally consider more complex forms of learning. They further invoke semantically rich norm representations and processes that come closest to what we can describe as *ideation* [Ehrlich and Levin, 2005], i.e. proposing behaviours that potentially qualify as normative, and selectively filtering those.

Representative works that apply cognitive approaches include the *Beliefs-Obligations-Intentions-Desires* (BOID) architecture [Broersen *et al.*, 2001; Broersen *et al.*, 2002] which extends the widely adopted Belief-Desire-Intention (BDI) architecture [Bratman, 1987; Rao and Georgeff, 1995] with an obligation component that preempts the goal generation and prioritizes the individuals' obligations. In this approach, obligations are statically embedded in an agent's belief base.

While BOID emphasizes normative reasoning, alternative approaches propose mechanisms to facilitate norm identification and decision-making, along with the involved micro-/macro-level interaction, as in the cognitive architecture EMIL [Andrighetto *et al.*, 2007; Campenni *et al.*, 2009; Andrighetto *et al.*, 2010], that extends the BDI concept with the ability to acquire new norms, which we discussed in Section 2.2.3.

Cognitive approaches such as Savarimuthu *et al.* 's norm identification frameworks for

[4] We will come back to this scenario in greater detail in Section 4, given of its relevance in the area of norm synthesis.

[5] We will discuss the field of norm emergence in more detail in Section 4.

obligation [Savarimuthu *et al.*, 2010b] and prohibition norms [Savarimuthu *et al.*, 2013a] rely on notions of machine learning to afford realistic agent representations [Savarimuthu *et al.*, 2011; Ossowski, 2013]. Further examples for the combined use of cognitive and machine learning components include the identification of normative content from action and/or event sequences (e.g. Savarimuthu *et al.* [2010a]), the implementation of alternative learning mechanisms beyond experiential learning or 'learning by doing', such as social/observational learning [Bandura, 1977] as applied by Epstein [2001], Hoffmann [2003], as well as Sen and Airiau [2007]. Another combined use of cognitive and machine learning is to facilitate the use of direct communication (e.g. used by Verhagen [2001] as well as Walker and Wooldridge [1995]).

Transmission The norm transmission process in Hollander and Wu's model (equivalent to the spreading process in Savarimuthu and Cranefield [2011]'s model), considers three components that characterize how information is spread. Those include:

- the nature of *Agent Relationships*,
- the applied *Transmission Techniques*, and
- the underlying *Network Structure*.

Agent Relationships Similar to Savarimuthu and Cranefield [2011], Hollander and Wu share Boyd and Richerson [1985]'s observation of relationship types as either being vertical, horizontal or oblique, an aspect we discussed in the context of Savarimuthu and Cranefield's model (Section 2.2.4).

Transmission Techniques Beyond the identification of relationships, Hollander and Wu [2011b] identify two transmission techniques for norms, the first being *active transmission* in which norms are actively broadcast throughout the relationship networks. Alternatively, agents can use *passive transmission* and absorb perceived normative information. Examples of mechanisms to facilitate active transmission include direct communication, whereas observation of the social environment (on the part of a norm recipient) is an example of passive transmission.

In most simulation works, active transmission is used to convey normative content by direct communication or in the form of sanctions. Examples include Hoffmann [2005], who uses proactively communicating norm entrepreneurs to promote convergence, as well as the previously mentioned work by Franks *et al.* [2013], or Yu *et al.* [2010; 2015]'s use of *supervisors* to model hierarchical communication between networked multi-agent systems. Further examples from the sociological domain include Castelfranchi *et al.* [1998]'s and Younger [2004]'s society models that rely on reputation sharing for the purpose of promoting cooperation.

Examples of passive communication are used to represent notions of imitation or social learning. Examples include Verhagen [2001]'s work on norms learning, as well as the work on the impact of social learning on norm convergence (e.g. Nakamaru and Levin [2004], Sen and Airiau [2007], Airiau *et al.* [2014]) and synthesis (e.g. Frantz *et al.* [2015]). An example of the use of passive transmission in social scenarios is Flentge *et al.* [2001]'s representation of imitation by copying memes from successful neighbours.

Network Structure The third aspect of norm transmission is the nature of the underlying connectivity structure that acts as an information transport medium. Depending on the objective, the connectivity structure is conceived as a multi-dimensional grid environment or as network topology of varying complexity.

In grid environments, agents are stationary or mobile, and observe agents within their specified neighbourhoods, and can, depending on their neighbourhood configuration, perceive adjacent cells. Agents' grid environments are generally modelled as von Neumann neighbourhoods – in which agents can sense orthogonally adjacent cells – or Moore neighbourhoods – in which agents can sense all adjacent cells.

The modelling of norm transmission via network structures permits the configuration of more complex relationship networks, with network topologies of equal degrees of connectedness (e.g. as fully connected networks), as well as random connectivity (random networks [Erdős and Rényi, 1959]). Alternatively, networks can display varying degrees of connectedness, such as small world networks [Watts and Strogatz, 1998] that simulate sparse links between communities characterized by dense internal connectedness. Scale-free network topologies [Barabási and Albert, 1999] work on the far end of the spectrum and produce a structure characterized by power law distributions, with individuals being centred around densely-connected hubs.

In analogy to the stationary or mobile configuration in a grid environment, a further important aspect is whether network topologies are static or dynamic at runtime. Effects of complex network topologies on norm emergence have been explored by Zhang and Leezer [2009], Franks *et al.* [2014], and Sen and Sen [2010]. Villatoro *et al.* [2009] put specific emphasis on the interaction between memory size and the chosen topology, whereas Airiau *et al.* [2014] concentrate on the effect of social learning across different topologies. Savarimuthu *et al.* [2007] and Villatoro *et al.* [2011a; 2013] explore the effect of dynamic topologies on norm emergence.

Recognition In Hollander and Wu's model, the processes *creation* and *transmission* are followed by the superprocess *enforcement* that consists of the subprocesses *recognition* and *enforcement* (see Figure 3). Norm recognition is similar to Savarimuthu and Cranefield's account of *norm identification* and describes the agent's ability to recognize the norms enacted in the observed society or group. Means to do so include communication with norm participants (as is the case with human societies [Henderson, 2005]) as well as observational learning. Similar to technological approaches in the context of norm creation, earlier models relied on off-line identification of agents as norm followers and deviants (e.g. Castelfranchi *et al.* [1998] , Hales [2002]), whereas recent models apply more sophisticated mechanisms to identify norms, which include machine learning [Sen and Airiau, 2007; Mukherjee *et al.*, 2007; Savarimuthu *et al.*, 2013b; Frantz *et al.*, 2015] and/or cognitive approaches [Savarimuthu *et al.*, 2010b; Andrighetto *et al.*, 2007]. Since the recognition of norms may involve the observation of sanctions, it is closely related to enforcement.

Enforcement Norm enforcement describes the application of sanctions to stimulate adherence to the normative content. Sanctions can be positive (in the form of rewards) or negative in nature and can further be differentiated by their source, that is whether they originate from internal or external sources.

For this purpose Hollander and Wu differentiate three types of enforcements:

- Externally Directed Enforcement
- Internally Directed Enforcement
- Motivational Enforcement

Externally Directed Enforcement Externally directed enforcement describes sanctioning by an outside observer that witnesses and reacts to a norm violation or an agent's refusal to accept a transmitted norm (e.g. a follower rejecting a leader's imposed norm) [Flentge *et al.*, 2001; Galan and Izquierdo, 2005; Savarimuthu *et al.*, 2008b].

Applied sanctions can be of economic nature (e.g. reducing or limiting access to resources), affect the violator's reputation (e.g. shunning, ostracism) [Axelrod, 1986; Castelfranchi *et al.*, 1998; Hales, 2002; Younger, 2004] (as seen in Section 2.2.5), or prevent it from propagating deviance to others (e.g. by preventing procreation in the case of vertical norm transmission [Flentge *et al.*, 2001]) [Caldas and Coelho, 1999].

The prototypical example for external sanctions is Axelrod's norm game [Axelrod, 1986], as discussed in Section 2.2.5 in the context of Savarimuthu and Cranefield's life cycle model.

Internally Directed Enforcement Sanctions of internal origin rely on an individual's self-enforcement triggered by the violation of internalized norms. The prototypical mechanism for internally motivated norm enforcement is the activation of emotions (discussed in greater detail in Section 2.2.5).

Motivational Enforcement Hollander and Wu further identify the notion of motivational enforcement, which is essentially a special case of internally directed enforcement. It describes the implicit commitment of all individuals to follow system-wide norms if they are aligned with an individual best interest, an aspect understood as conventions [Lewis, 1969]. A classical example is the convention of uniform road side use: the precise strategy (i.e. whether to drive on the left or right side) is secondary to the complete acceptance and internalization by the society since unilateral deviation produces suboptimal outcomes (i.e. accidents caused by ghost drivers).

Internalization Processes that are essential for norm emergence in Hollander and Wu's model are associated with the superprocess norm internalization. Hollander and Wu differentiate between *Acceptance*, *Modification*, and *Internalization* (as the terminating subprocess of the superprocess *Internalization*).

The acceptance of enforced norms is the starting point for the internalization of norms by individuals and decisive for the emergence of norms, since individuals either decide to accept or reject socially imposed norms based on the compatibility with their personal beliefs, desires and intentions. Possible outcomes are the acceptance of a new norm, the substitution of an existing conflicting norm, or its rejection. Acceptance is operationalized as some form of cost-benefit analysis [Meneguzzi and Luck, 2009].

If agents decide to accept norms, their integration into the internal cognitive structures requires the transformation of norms from an objectified outside perspective to a subjective representation that involves an individual's biases, inaccuracies of perception, etc. This potentially leads to a modified understanding of that norm, an aspect that affects the norm during its further progression in the life cycle.

Finally, the accepted and potentially modified norm is internalized by the receiving agent. Compared to the other stages of the norm life cycle, this process has found limited explicit attention. In most applications, individuals simply adopt the accepted norms without further refinement or adaptation. From a motivational perspective, this is compatible with measures that suggest that the absence of external pressures is indicative of complete norm internalization [Epstein, 2001]. However, this view only accounts for subsequent norm adherence, but cannot explain violations further down the track. Refined approaches evaluate the effect of the internalized norm and on the decision-making process. An important example is Verhagen [2001]'s work, in which agents seek increasing alignment with their associated group by comparing and internalizing corresponding action probabilities. Alternatively, as done in the BOID architecture [Broersen *et al.*, 2001], internalized norms can be maintained separately from personal strategies and activated selectively depending on situation-specific autonomy values [Broersen *et al.*, 2002].

In their original survey, Hollander and Wu [2011b] highlighted the limited explicit focus on internalization, especially in comparison to life cycle processes such as enforcement. However, recent works in the area of NorMAS reveal more explicit treatments of internalization, generally in the form of continuous probabilistic adaptation of strategy choices based on reinforcement learning (e.g. Salazar *et al.* [2010], Villatoro *et al.* [2013], Franks *et al.* [2014], Airiau *et al.* [2014], Frantz *et al.* [2014b; 2015], Yu *et al.* [2015]), or by using thresholds for the adoption of new strategies (e.g. Hollander and Wu [2011a], Mihaylov *et al.* [2014]). In Section 2.6 we provide a comprehensive overview of internalization mechanisms used in works on normative multi-agent systems.

Emergence In contrast to all earlier models, Hollander and Wu conceive emergence as a dynamic macro-level process that describes a cyclic iteration involving the transmission of the internalized norm to new participants. This is followed by enforcement (based on the subprocesses *Recognition* and *Enforcement*) to drive the internalization (composed of subprocesses *Acceptance*, potential *Modification*, and *Internalization*) of the norm by new subjects, who themselves participate in the spreading of the norm – ultimately leading to the norm's emergence as a macro-level phenomenon. This emergence understanding is aligned with Savarimuthu and Cranefield's, who interpret emergence as the final stage of the norm life cycle, but do not explicitly reflect the cyclic reinforcement of norms by reiterating through the formation stage. Finnemore and Sikkink's life cycle model maintains a different emergence interpretation and associates emergence with the micro-level creation of a norm, e.g. via entrepreneurship, before sharing and penetrating the wider society.

The exploration of emergence characteristics is strongly tied to the applied modelling technique. Game-theoretical approaches evaluate emergence by identifying stabilising strategy choices (equilibria) chosen from a set of given alternative strategies. The dominant strategy choice is then interpreted as the emergent norm (see e.g. Axelrod [1986], Mukherjee *et al.* [2007], Zhang and Leezer [2009]). Since agents are represented as structurally uniform selfish rationalizers with a minimal action repertoire, the exploration is focused on macro-level outcomes. Cognitive approaches, on the other hand, do permit a macro-level observation of specific norms, but furthermore, allow a more realistic reconstruction of micro-level processes. This includes detail and diversity of individuals' cognitive structures, the precise level and nature of enforcement (see e.g. Caldas and Coelho [1999], Savarimuthu *et al.* [2008b]), the use of richer norm representations, diverse action sets, and a variety of norm learning mechanisms (e.g. based on experiential learning, social learning and direct communication) [Savarimuthu *et al.*, 2011].

Models can further address infrastructural aspects, such as the impact of different connectivity structures on normative outcomes. Related findings suggest that scenarios in which normative behaviour is transmitted from neighbours (e.g. in grid environments) tend to result in the dominance of a single norm, whereas individualized learning promotes the emergence of diverse normative configurations [Boyd and Richerson, 1985; Boyd and Richerson, 2005; Nakamaru and Levin, 2004]. While the application of network structures can lead to stronger normative diversity, experimental results suggest that the impact of the actual network topology is secondary to its dynamic nature (as opposed to static networks) [Bravo *et al.*, 2012]. However, the convergence of conventions (and emergence of local subconventions) can be controlled by maintaining links to distant nodes [Villatoro *et al.*, 2009].

Forgetting & Evolution In contrast to the earlier models by Finnemore and Sikkink as well as Savarimuthu and Cranefield, Hollander and Wu are the first to complete the norm life cycle by explicitly considering the process of *Forgetting*. In this conception forgetting is essential to sponsor the evolutionary refinement of norms, since continuously changing norm contexts may render existing norms irrelevant. An example is the normalized use of smart devices in education, with proactive integration of social media platforms such as Facebook into the learning environment. This is in opposition, or at least in competition, to traditional norms that ban the use of mobile devices in classroom environments. Once forgotten, norms make space for new norms that are better adapted to environmental needs, which constitutes the end-to-end process that closes the evolutionary loop of the norm life cycle.

2.3.3 Discussion

As mentioned at the outset of this section, this model proposed by Hollander and Wu introduces the to date most comprehensive life cycle model. The model not only considers abstract high-level processes (superprocesses), but decomposes those into elementary processes that capture large parts of contemporary research and, beyond this, identify gaps in normative agent architectures (such as the explicit consideration of *Norm Acceptance*) to produce more comprehensive representations of human reasoning processes. In addition to the fine-grained nature, this model further deviates from the linear operation of previous models by identifying emergence as a metaprocess that links individual processes and results in a continuous iteration through elementary processes. Beyond the ‘completion’ of the life cycle by considering the abandoning of norms, a further essential novelty is the consideration of norm evolution as a continuous process that affords both the modification and the substitution of norms over time.

2.4 Model 4: Mahmoud et al.

Overview The latest life cycle model has been proposed by Mahmoud *et al.* [2014b]. Similar to the earlier life cycle models developed in the context of NorMAS, their work is based on a comprehensive literature review, both considering individual works as well as previous life cycle models. In contrast to Hollander and Wu's detailed model, their approach identifies five core processes (*Creation*, *Emergence*, *Assimilation*, *Internalization*, *Removal*) with a further decomposition of selected processes as shown in Figure 4. Since this model has only been briefly described by the original authors themselves and strongly builds on concepts introduced in the context of Hollander and Wu's earlier, more detailed model, we provide a concise overview at this stage, before discussing the novel contributions in more detail.

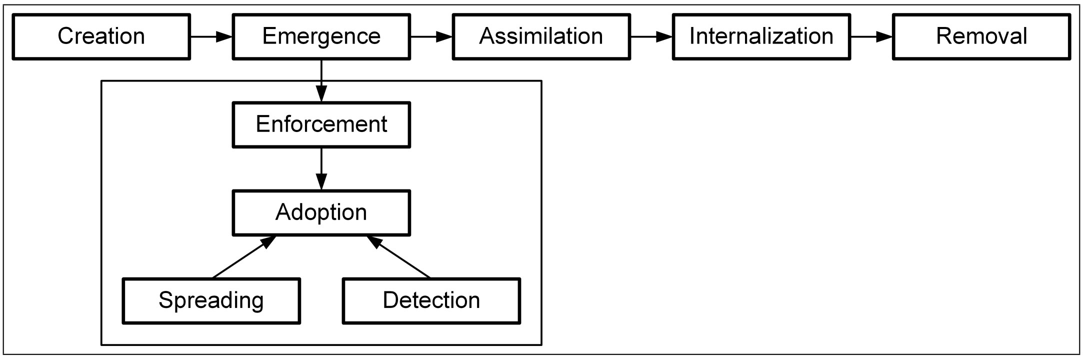

Figure 4: Mahmoud et al.'s Norm Life Cycle Model

Processes The initial process, as with most other life cycle models is *Creation*, which operates based on mechanisms described by Savarimuthu and Cranefield [2011], namely off-line design, autonomous innovation and social power (see Section 2.2).

A central deviation from previous models is the process of *Emergence*, which Mahmoud *et al.* decompose into two individual processes, *Norm Enforcement* and *Norm Adoption*. The latter of those is further decomposed into the processes *Norm Detection* and *Norm Spreading*. Unlike Hollander and Wu's model, emergence is considered a sequential process.

In Mahmoud *et al.* 's model, *Enforcement* consists of direct and indirect sanctioning, where direct sanctioning is the conventional application of reward or punishment, whereas indirect sanctioning is reflected in an individual's reputation and emotions (e.g. guilt).

The *Adoption* process is a composite process that consists of the spreading of new norms and the detection of norms. The *Spreading* process captures the transmission directions outlined by Savarimuthu and Cranefield [2011] (vertical, horizontal and oblique). The *Detection* of new norms captures all forms of norm learning to identify new norms, including imitation, social learning, case-based reasoning and data mining. The model further emphasizes the essential nature of network topologies to facilitate the spreading of norms, including the differentiation of static and dynamic networks, but does not consider alternative mechanisms such as sensing in grid-based environments.

Following the *Emergence* process, the model introduces a novel *Assimilation* process. The authors follow Eguia [2011]'s definition of assimilation "as the process in which agents embrace new social norms, habits, and customs, which is costly but offers greater opportunities" ([Mahmoud *et al.*, 2014b], p.15). In their conception, assimilation involves deciding whether to adopt new social norms by trading off associated costs and benefits.

This process is followed by the *Internalization* process that, similar to Hollander and Wu's conception, includes the *Acceptance*, *Transcription* and *Reinforcement* of the newly acquired norm, with the purpose of embedding it in the agent's behaviour.

The final *Removal* process is equivalent to Hollander and Wu [2011b]'s process of forgetting norms. The purpose is the removal of obsolete norms, as well as being an implicit consequence of norm modification. Mahmoud *et al.* further adopt an unspecified end-to-end process that links *Removal* and *Creation*, possibly implying the evolutionary process introduced by Hollander and Wu.

Discussion The model by Mahmoud *et al.* breaks the trend of proposing progressively more detailed models and attempts to identify the essential processes instead. This condensed conception produces an incoherent understanding of the norm life cycle and semantic ambiguities, the causes of which we will explore in the following section.

Despite the authors' awareness of previous models, in this model emergence only considers the enforcement and adoption of norms (which captures aspects such as spreading and detection), but does not consider the internalization of norms essential for their emergence. How norms can emerge without being internalized is left unexplained. This leaves unclear whether internalization is implied as part of the *Adoption* process that concentrates on spreading and detection of norms. If this were the case, this would produce an ambiguous understanding of the subsequent internalization process.

A similar problem relates to the novel *Assimilation* process, which represents the authors' own substantive contribution [Mahmoud *et al.*, 2014a] to the field of NorMAS. Since assimilation describes the process of deciding whether to adopt given norms, it is unclear in how far this is different from the *Acceptance* process that is part of norm internalization [Mahmoud *et al.*, 2014b], or if it is meant to replace the acceptance component of internalization. The authors' related contribution [Mahmoud *et al.*, 2014a] discusses the assimilation of norms in heterogeneous communities and suggests that the norm internalization itself is a *subprocess* of norm assimilation, an aspect that is not reflected in the sequential organisation of both processes in the life cycle model (see Figure 4). The inspection of the authors' related work suggests that assimilation not so much describes a norm-centred life cycle process. Instead, it characterizes an agent's capability since it describes the *ability and willingness of agents to integrate into their social environment* [Mahmoud *et al.*, 2014a], which entails the adoption of norms, customs, habits, etc.

Overall, the model attempts to rationalize the existing norm life cycle models, leading to a refined but insufficiently specified and contextualized life cycle model, specifically with respect to the emergence process as well as the novel assimilation component – aspects that challenge its coherence and, in consequence, applicability.

2.5 Comprehensive Literature Overview

In the previous sections, we introduced the most relevant life cycle models known in the literature and discussed associated significant contributions. Table 1 integrates the mentioned literature into a comprehensive chronological overview that spans across selected life cycle processes.[6] Whereas the process characteristics of creation, identification, spreading, and enforcement are based on the criteria and approaches discussed in the context of the individual life cycle models (specifically in Sections 2.2 and 2.3), this overview puts particular focus on capturing internalization mechanisms and emergence characteristics, both of which have found limited recognition in previous surveys.

Earlier works on norm internalization apply the specification of norms at design time, which occurs in conjunction with off-line norm creation (which we labelled 'embedded'). However, in the majority of contributions, the adoption and internalization of norms generally occur unreflected (labelled 'immediate'). In more recent approaches, we can observe a shift towards more continuous internalization of norms based on observation ('social learning') as well as probabilistic or threshold-based adoption based on sustained reinforcement ('threshold-based learning', 'Q-learning').

Another category that is characterized by a range of varying, often scenario-dependent measures is the notion of emergence. Examples include convergence thresholds on shared equilibrium strategies in the case of coordination games. In alternative approaches emergence refers to the alignment of sets of norms, both including crisp (e.g. Campenni *et al.* [2009], Andrighetto *et al.* [2010], Griffiths and Luck [2010]) and fuzzy set conceptions (e.g. Frantz *et al.* [2014b; 2016]), or the identification of a shared normative understanding, e.g. by election (Riveret *et al.* [2014]) or by generalization (Frantz *et al.* [2015]). Another group of approaches interpret emergence as the convergence on shared conceptualisations of lexica (e.g. Salazar *et al.* [2010], Franks *et al.* [2013]).

[6] This overview refines and extends an earlier survey produced by Savarimuthu and Cranefield [2011].

Publication	Creation	Identification	Spreading	Enforcement	Internalization	Emergence
Axelrod [1986]	-	-	vertical	Sanctions	immediate	Converging strategy choice
Kittock [1995]	-	Machine learning	-	Sanctions	memorizing strategy	Converging strategy choice
Conte and Castelfranchi [1995b]	Off-line design	-	-	-	embedded	Survival under different strategies
Walker and Wooldridge [1995]	-	Machine learning	-	-	-	-
Shoham and Tennenholtz [1992b; 1995]	Off-line design	Machine learning	-	-	-	-
Shoham and Tennenholtz [1997]	-	-	-	Sanctions	immediate	Converging strategy choice
Castelfranchi *et al.* [1998]	Off-line design	-	-	Reputation	embedded	-
Saam and Harrer [1999]	Off-line design	-	oblique	Sanctions	embedded	-
Verhagen [2001]	Leadership	-	oblique	Leader/group feedback	alignment with group	Social alignment of action probabilities
Epstein [2001]	-	Machine learning	horizontal	-	imitation	Converging on action choice
Flentge *et al.* [2001]	-	-	vertical	Sanctions	inherited	-
Hales [2002]	Off-line design	-	-	Reputation	immediate	-
Hoffmann [2003]	Entrepreneurship	-	oblique	Reward	immediate	Converging on chosen value
Delgado [2002; 2003]	-	Machine learning	horizontal	Payoff	immediate	Converging on state
López y López and Luck [2004]	Off-line design	-	-	Sanction/Reward	-	-
Nakamaru and Levin [2004]	-	Machine learning	horizontal	-	immediate	Stabilising norms
Pujol *et al.* [2005]	-	Machine learning	-	-	-	Converging strategy choice
Chalub *et al.* [2006]	-	Machine learning	vertical	Reputation	immediate	Converging norms
Fix *et al.* [2006]	-	-	-	Emotion	immediate	-
López y López *et al.* [2006; 2007]	Off-line design	-	-	Sanction/Reward	-	-
Sen and Airiau [2007]	-	Machine learning	-	-	immediate	Converging strategy choice
Mukherjee *et al.* [2007; 2008]	-	Machine learning	-	Payoff	immediate	Converging strategy choice
Savarimuthu *et al.* [2007; 2008a]	-	Machine learning	-	-	Learning	Converging on value
Campenni *et al.* [2009; 2010]	-	Cognition, social learning	-	Payoff	-	Shared event-action trees
Savarimuthu *et al.* [2009]	-	-	oblique	-	immediate	Converging on shared norm
Urbano *et al.* [2009]	-	-	oblique	Payoff	immediate	Converging towards joint strategy
Villatoro *et al.* [2009]	-	Machine learning	-	Reward	immediate	Converging action choice
Sen and Sen [2010]	-	Machine learning	horizontal	-	immediate	Converging towards joint action
Griffiths and Luck [2010]	-	-	vertical	-	immediate	Convergence on multiple norms
Savarimuthu *et al.* [2010b; 2010a; 2011]	Off-line design	Cognition, data mining	-	Sanction signal	immediate	Identification of event sequences as norms
Perreau de Pinninck *et al.* [2010]	-	-	-	Ostracism	-	Minimal norm violations
Salazar *et al.* [2010]	-	-	horizontal	-	probabilistic	Convergence on shared word/concept lexicon
Yu *et al.* [2010]	-	Machine learning	oblique	-	social learning	Convergence on shared goal state
Sugawara [2011]	-	Machine learning	horizontal	Payoff	Q-learning	Convergence on non-conflicting conventions
Villatoro *et al.* [2011b]	-	-	horizontal	Payoff	social learning	-
Hollander and Wu [2011a]	-	Machine learning	horizontal	internal	threshold-based learning	Equilibrium strategies (various scopes)
Mahmoud *et al.* [2012; 2015]	-	-	oblique	dynamic sanctions	-	Converging action choice
Riveret *et al.* [2012; 2013]	-	Machine learning	horizontal	Sanction	Learning	-
Savarimuthu *et al.* [2013b]	-	Machine learning	-	-	-	-
Villatoro *et al.* [2013]	-	Machine learning	horizontal	Payoff	Q-learning	Converging action choice
Franks *et al.* [2013]	-	Evolutionary algorithm	oblique	Leadership	probabilistic	Convergence on shared word/concept lexicon
Morales *et al.* [2013; 2014; 2015b]	-	Machine learning	oblique	Sanction	immediate	Synthesised set of norms
Mihaylov *et al.* [2014]	-	Machine learning	horizontal	Payoff	threshold-based learning	Converging action choice
Airiau *et al.* [2014]	-	Machine learning	horizontal	-	Q-learning	Converging strategy choice
Frantz *et al.* [2014b; 2016]	-	Machine learning	-	Payoff	Q-learning	Alignment of fuzzy normative understanding
Riveret *et al.* [2014]	-	Machine learning	oblique	Payoff	Q-learning	Shared normative understanding
Frantz *et al.* [2014c; 2015]	-	Machine learning	horizontal	Sanction/Reward	Q-learning	Multi-level norm generalization
Yu *et al.* [2015]	-	Machine learning	oblique	-	Q-learning	Converging strategy choice
Beheshti *et al.* [2015]	-	Cognition	horizontal	Sanction/Reward	Q-learning	Norm-conforming behaviour

Table 1: Chronological Overview of Literature and Associated Life Cycle Characteristics

2.6 Systematic Comparison of Norm Life Cycle Models

To this stage, we have introduced a diverse set of life cycle models along with associated contributions, but have yet to relate those systematically. Finnemore and Sikkink [1998] 's model (Section 2.1), proposed in the field of international relations, identifies three processes in a norm's life, starting with its explicit creation (*Emergence*), its spreading (*Cascade*), leading to wide-ranging adoption (*Internalization*). In contrast to all other models, their model looks at states as central players and emphasizes the long-term perspective of normative change (e.g. embedding the changing societal normative view in professional ethics).

The remaining three ones are products of systematic reviews of contemporary research in the area of NorMAS, an approach spearheaded by Savarimuthu and Cranefield [2011] . Their model (Section 2.2) provides a refined account of the beginning of a norm's development, with a particular focus on the initial formation and propagation. Their model interprets emergence as an outcome measure and does not include a long-term perspective on norms, such as their decay and substitution over time.[7] However, since their model is grounded in a systematic review of existing works, this does not indicate a principle shortcoming of the model, but rather reflects the contemporary state of the research field.

Hollander and Wu [2011b]'s model (Section 2.3) provides the most comprehensive account of norms' life cycles, and, similar to Savarimuthu and Cranefield's grouping of processes into stages, identifies essential superprocesses that are composed of refined subprocesses. Their model goes beyond previous accounts and proposes processes that are only weakly reflected in literature, thus identifying presumed research gaps. The most important contribution of their model is the recognition of cycles of recurring processes, an example of which is the characterisation of norm emergence as a reiteration of transmission, enforcement and internalization. The second essential contribution is the integration of a long-term perspective on normative change, which they reflect as an evolution process.

Finally, Mahmoud *et al.* [2014b] (Section 2.4) describe a model that condenses the number of relevant processes of the normative life cycle to five. Their model puts specific emphasis on norm assimilation, i.e. an individual's decision whether to accept (and subsequently internalize) a given norm. They further decompose the emergence process into enforcement and adoption (which in itself consists of the processes *Norm Spreading* and *Norm Detection*). Similar to Finnemore and Sikkink, as well as Savarimuthu and Cranefield, Mahmoud *et al.* conceive a linear norm life cycle; they do not consider iterative processes.

An aspect that challenges the systematic comparison of all four models is not only the varying level of detail, but the observable terminological ambiguity. In the different life cycle models the sharing or spreading of norms is selectively captured by the terms 'cascade' (Finnemore and Sikkink), 'spreading' (Savarimuthu and Cranefield, Mahmoud *et al.*), and

[7] They consider those as part of a refined set of life stages in later work [Savarimuthu *et al.*, 2013b].

‘transmission’ (Hollander and Wu). A further notable example is the norm ‘identification’ (Savarimuthu and Cranefield) that is alternatively characterized as ‘recognition’ (Hollander and Wu) or ‘detection’ (Mahmoud *et al.*).

Beyond those synonyms, the specific processes in different models have semantic overlappings. To facilitate a systematic comparison of content and semantic relationship, in Figure 5 we provide an overview of all life cycle models, with individual processes roughly aligned by semantic relationship. Process labels are formatted and grouped to reflected their nature and importance in the respective life cycle model:

- Savarimuthu and Cranefield differentiate between individual processes and stages. Consequently, the *life cycle stage names* are held in bold font.
- Hollander and Wu ’s *superprocess labels* are held in bold font. The emergence and evolution processes are further explicitly included in the schematic overview.
- Mahmoud *et al.* ’s model composes the emergence process from two elementary processes and is thus held in bold font, along with all further *processes of the same conceptual weight*.

Dotted lines indicate the semantic relationships between individual processes of the corresponding life cycle models. For example, Finnemore and Sikkink ’s *Cascade* process combines components of Savarimuthu and Cranefield ’s *Spreading* and *Enforcement* processes.

Despite the diversity of norm life cycles, the systematic review of all models reveals clusters of processes that have similar or identical functions (identified as solid horizontal lines in Figure 5). We can generalize four such clusters, or phases, of norm life cycles, and label those by complementing the labels of the initial two life cycle stages in Savarimuthu and Cranefield [2011]’s model:

- Formation – Processes associated with the creation and inference of norms
- Propagation – Processes associated with the communication of norms
- Manifestation – Processes associated with the general acceptance and entrenchment of norms
- Evolution – Processes associated with the evolutionary refinement of norms

The identified phases correspond to the abstract phases proposed by Andrighetto *et al.* [2013], namely *Generation*, *Spreading*, *Stability* and *Evolution*, an aspect that supports the semantic process clusters proposed above. Our terminological choice is driven by the goal

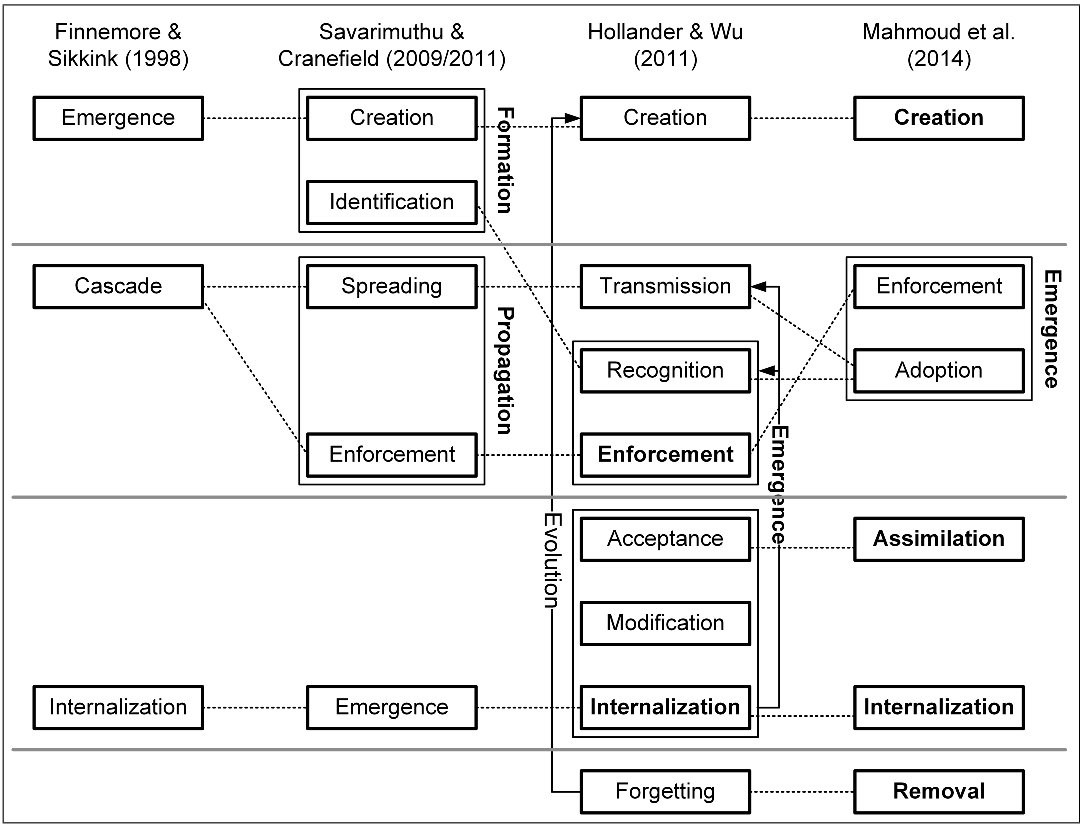

Figure 5: Schematic Comparison of Discussed Norm Life Cycle Models

to comprehensively capture the semantics of associated processes of all discussed norm life cycle models (e.g. operations associated with norm internalization extend beyond the characterization of a norm as stable – see discussion below). In the following, we will use the identified phases to compare and contextualize the norm life cycle models.

Phase 1: Formation All models identify norm creation as the initial life cycle step. In contrast to all other models, Finnemore and Sikkink [1998] employ a different emergence understanding. In their conception emergence entails the initial creation of a norm (which Hollander and Wu [2011b] describe as "norm creation on a micro scale" [Hollander and Wu, 2011b]), whereas life cycle models from the area of NorMAS (henceforth referred to as NorMAS models) understand emergence as "norm establishment on a macro scale" [Hollander and Wu, 2011b]. However, the underlying understanding of this initial phase – the explicit creation of a norm – is identical for all models. Despite this uniform characterization, we label this phase as *Formation* in order to capture a more general understanding of norm creation, widening the scope to approaches that do not rely on explicit norm creation such as the identification of existing/unknown norms by observation, an aspect implicitly captured by Savarimuthu and Cranefield's notion of *Norm Identification* (which we discuss in Section 4).

Phase 2: Propagation Following the creation, all models describe some sort of norm communication, or propagation (*Cascade*, *Spreading*, *Transmission*, and *Adoption*). A special case is Mahmoud *et al.* [2014b]'s *Adoption* process, which entails both norm spreading and detection. All NorMAS models recognize a notion of norm identification (*Identification*, *Recognition*, and *Adoption*), but have a varying sequential organisation. While Savarimuthu and Cranefield's early allocation of norm identification is driven by the understanding that agents need to identify norms in their environment, all subsequent models interpret it as a step that follows the transmission of a norm. Similarly, all NorMAS models recognize enforcement as an essential determinant of a norm's success.

Phase 3: Manifestation The propagation of norms is followed by their *Internalization*. In Finnemore and Sikkink's model that refers to the wide-ranging adoption of a norm within society and its embedding in societal institutional structures. In addition to gaining stability, at this stage norms thus manifest themselves in the social fabric which implicitly reinforces their persistence, constrains future action, but also limits the potential of competing norms. Manifested norms can attain quasi-legal status, e.g. by shaping the codes of ethics for specific occupations, which are subsequently absorbed into the discipline's professional training and practices. This understanding is compatible with Savarimuthu and Cranefield's *Emergence* interpretation, which represents the extent to which a norm is able to penetrate the affected society.

While these first two models describe norm manifestation as a macro-level process, the models of Hollander and Wu as well as Mahmoud *et al.* describe refined sets of micro-level processes that lead to the internalization of norms. Hollander and Wu differentiate between *Acceptance*, *Modification*, and *Internalization*, including the decision whether to adopt a norm in the first place, and reflecting individual biases introduced during internalization. Mahmoud *et al.* reduce those to two processes, namely *Assimilation* and *Internalization*. As discussed in Section 2.4, the authors borrow the notion of *Acceptance* (which is identical to Hollander and Wu's *Acceptance*[8]), and consider it part of the *Internalization* process.

[8] "Norm acceptance is a conflict resolution process in which external social enforcements compete against the internal desires and motivations of the agent. If the new norm is in conflict with existing norms and may lead to inconsistent behaviours, or if the cost of accepting the new norm is too high, it will be rejected ..." [Hollander and Wu, 2011b], paragraph 3.24.

However, they introduce a preceding *Assimilation* process[9] (whose function is not clear, since it is not sufficiently contrasted to *Acceptance*) and *Internalization*. At the end of this manifestation phase, all models assume that individuals have embraced the promoted norms.

Phase 4: Evolution The fourth phase which we tag *Evolutionary Phase* is only reflected in the later life cycle models which introduce the processes *Forgetting* and *Removal* that reflect the end of the normative life cycle. However, more important than their function to 'complete' the norm life cycle is their role as starting point for an evolutionary process (as introduced by Hollander and Wu [2011b]; Section 2.3) in which norms are refined or substituted by more relevant or efficient norms; forgetting old norms is a by-product of this evolutionary refinement and technical necessity to maintain efficient but also realistic architecture implementations.

The 'Special Case' Emergence Only exception to the uniform organisation of processes into general phases is the notion of emergence, which reflects the terminological ambiguity surrounding this concept. Whereas Finnemore and Sikkink's micro-level interpretation of emergence is associated with the *Formation Phase*, Mahmoud *et al.* see the *Propagation Phase* with the processes of *Enforcement* and *Adoption* as decisive for emergence. Hollander and Wu see emergence as an iterative process that spans across *Formation* and *Manifestation Phase*. Savarimuthu and Cranefield associate emergence with the third phase of norm manifestation and interpret it as a result of *Formation* and *Propagation*.

We believe that Hollander and Wu's cyclic representation represents the most accurate characterisation of the emergence process, since it links the macro-level emergence process with the underlying propagation and internalization processes, an aspect we will revisit in the context of proposed refinements (see Section 2.7). Savarimuthu and Cranefield's interpretation as outcome measure only reflects a quantifiable macro-level phenomenon, but does not maintain its relationship to the underlying processes that produce it. Mahmoud *et al.* inherently rely on propagation processes to determine a norm's emergence. Their model neither considers the cyclic nature of emergence nor does it consider the internalization of norms as a precursor for their further spread (see discussion in Section 2.4).

Norm Life Cycle Models and Levels of Analysis Comparing the individual models leaves the general impression that later models (with exception of Mahmoud *et al.*) are increasingly detailed and comprehensive. However, while this observation is warranted, it rather reflects the operational levels the life cycle models represent. Finnemore and Sikkink's,

[9] "[Norm assimilation is] ... the process in which agents embrace new social norms, habits and customs, which is costly but offers greater opportunities." [Mahmoud *et al.*, 2014b], p.15 with reference to Eguia [2011].

as well as Savarimuthu and Cranefield's models, describe the adoption and implementation of norms on the macro level, i.e. group or society level. This is well captured in Finnemore and Sikkink's understanding of internalization as the process of embedding the norm in a society's social structures and institutions. Similarly, Savarimuthu and Cranefield describe emergence as a macro-level outcome that describes the adoption of a norm across the wider society. Hollander and Wu's model introduces a shift from the macro-level norm perspective to an individual-centred micro-perspective, an aspect that is particularly apparent in the elementary processes they describe in the context of the establishment phase. Micro-level processes include *Acceptance* (the decision whether or not to accept norms), *Modification* (the modification of norms during internalization based on individual biases), and finally *Internalization*, which describes an individual's integration of norms into its existing belief structure. Only the subsequent *Emergence* and *Evolution* processes operate on the macro level, since they shift the perspective from individual to society level. Mahmoud *et al.* 's model similarly emphasizes individual-level processes such as *Assimilation* and *Internalization*, which they decompose into operational steps that are similar to Hollander and Wu's processes (Mahmoud *et al.*: *Acceptance*, *Transcription*, *Reinforcement*; Hollander and Wu: *Acceptance*, *Modification*, *Internalization*). In both models forgetting and removal of norms emphasizes a micro-level operation and is considered a technological necessity (in the light of limited computational resources), but obscures the macro-level function of facilitating an evolutionary refinement [Hollander and Wu, 2011b] of the normative landscape.

Understanding the different operation levels of the introduced models is helpful, since it allows their selective consultation. For the modelling and analysis of macro-level phenomena, the use of Savarimuthu and Cranefield's model may provide sufficient conceptual backdrop, whereas detailed cognitive agent models will find the most comprehensive structural blueprint in Hollander and Wu's model, with other models providing even higher levels of abstraction (Finnemore and Sikkink) or varying emphasis of individual-level processes (Mahmoud *et al.*).

2.7 General Norm Life Cycle Model

As a result of reviewing the existing life cycle models and their respective biases, we propose a general life cycle model that harmonizes various inconsistencies of the introduced approaches (e.g. micro- vs. macro-level operation, emergence understanding), but also addresses explicit conceptual omissions that are of increasing importance in recent developments (see Sections 3 and 4).

As such, the proposed general norm life cycle model introduces five essential revisions, which we discuss in the following:

- Distinction between micro-level processes and macro-level phenomena

- Norm Identification as an alternative life cycle entry point (in addition to explicit norm creation)
- Enforcement as a dynamic process with norm emergence as a resulting phenomenon
- Norm Forgetting as by-product of norm evolution
- Potential norm modification throughout all life cycle processes

Distinction between Micro-Level Processes and Macro-Level Phenomena As discussed in great detail in the previous Section 2.6, the existing norm life cycle models operate on varying levels of abstraction, with the initial models identifying coarsely-structured processes, whereas the latter two models describe processes of varying granularity (e.g. Hollander and Wu's end-to-end processes, superprocesses in addition to regular processes). We propose a systematic distinction by separating the micro-level processes (e.g. Transmission, Identification and Internalization) that find explicit representation in normative architectures, from macro-level phenomena that arise from the cyclic operation of the underlying processes. We believe that differentiating between a processual and phenomenological perspective on norms is useful to inform modelling considerations in different problem domains, such as the engineering of a process-centric normative agent architecture, in contrast to macro-level processes such as the emergence of norms within agent societies or their evolution over time. However, at the same time, these perspectives should not be dissociated in order to retain the links between the phenomena and the underlying processes. *Norm Emergence* is thus a result of iterative Transmission, Identification, Internalization and Enforcement processes. *Norm Evolution* extends across the entire norm life cycle, additionally involving the inception of new norms (Norm Creation) as well as the forgetting of decaying norms (Norm Forgetting).

Norm Identification as a Life Cycle Entry Point To date, the existing approaches assume the explicit creation of a norm. Proposed mechanisms include norm leadership, entrepreneurship, autonomous innovation and social power. However, in reality, norms may not necessarily be explicitly created, of unknown origin, but be rooted from behavioural regularities based on individuals' necessity to act in the first place (described as "urgency of practice" [Bourdieu, 1977]). In principle, a situational strategy choice to coordinate behaviour (e.g. chosen means of greeting, road-side choice) can emerge as self-enforcing convention (without intentional explicit conceptualisation), before finding recognition as a fully fledged norm.[10] Previous works acknowledge the existence of natural emergence pro-

[10] Examples for works that showcase this characteristic (e.g. Morales *et al.* [2015a], Riveret *et al.* [2014], Frantz *et al.* [2015]) are discussed in the context of the upcoming Section 4.

cesses[11] (Boella *et al.* [2008], Finnemore and Sikkink [1998], López y López *et al.* [2007], Savarimuthu and Cranefield [2009]), but assume an explicit creation as the starting point of the normative life cycle. We propose that a comprehensive norm life cycle should reflect the unplanned inception of norms based on social interaction as a possible alternative starting point of a norm's life – in addition to the explicit creation.

Enforcement as a Dynamic Process A further aspect relates to the role of enforcement. All NorMAS life cycle models represent enforcement as an explicit process that appears independent of notions such as spreading. However, enforcement itself can be interpreted as a dynamic process that promotes the cyclic reinforcement of norms, leading to their spread and thus their increasing adoption, producing emergence as an associated phenomenon (as discussed in the previous paragraph). Some form of enforcement – whether implicitly (e.g. serving as a guiding role model or influence based on shared values) or explicitly (e.g. by engaging in overt sanctioning) – is a prerequisite for the transmission of norms. In this context, it is further important to note that enforcement does not carry a specific valence, but can bear positive associations, such as providing a reward for a norm-compliant employee, or represent an explicit punishment, such as humiliating an individual in front of her reference group (e.g. an employee amongst fellow co-workers). Apart from such forms of overt *external enforcement*, enforcement can further be directed at oneself (internal enforcement), reflected in emotions such as the "warm glow" [Andreoni, 1989] of compliance (i.e. 'doing the right thing') or the guilt of violation (e.g. engaging in jaywalking despite conventional compliance).

Whether implicit or explicit, positive or negative, internal or external, enforcement relies on the prior internalization by the potential enforcer. This does not necessarily imply that the enforcer applies this norm to her- or himself or even 'believes' in it. As such, individuals can be tasked with the enforcement or feel pressured to defend norms they object to (such as not engaging in jaywalking in the presence of bystanders). Similarly, not violating a norm when facing the opportunity (without actively promoting it) can act as norm reinforcement. An example for this is the rejection of a bribe, especially if the actor holds a role model function (e.g. as a manager) [Hogg, 2001]. Conversely, the observation of violation by an authority figure (e.g. taking a bribe) can accelerate norm erosion. Whether compliant or not, essential for any positive or negative enforcement is some internalized conceptualization of the enforced norm in order to make its compliance and violation detectable. Consequently, we do not see emergence as a process in itself, but as a phenomenon that results from a sustained cyclic reinforcement based on the transmission, identification, internalization, and subsequent enforcement of norms, leading to their manifestation.

[11] Here emergence should be understood as the micro-level process of norm inception.

Forgetting as a By-Product of Norm Evolution A final aspect relates to the notion of forgetting. Hollander and Wu introduce forgetting as an end point of an evolutionary cycle that affords a norms refinement. However, the conceptualisation as an 'end-to-end process' presents it as a sequential step in a series of processes. Similar to the conception of emergence laid out before, we see evolution as a phenomenon that arises from the continuous reinforcement of norms, their change during identification and internalization, as well as their potential to become obsolete and ultimately forgotten. This process cannot be conceived as sequential but operates concurrently, with newly identified norms gaining more salience and potentially leading to existing norms' adaptation or decay. Though forgetting is an essential endpoint in the normative life cycle, it does not represent the starting point for a continuously operating evolution process; 'forgetting' is a by-product of evolving norms.

Schematic Overview In Figure 6 we show a schematic overview of the proposed refined norm life cycle that condenses elements of the previously introduced models, but incorporates essential revisions. We will briefly explore the processes in the following.

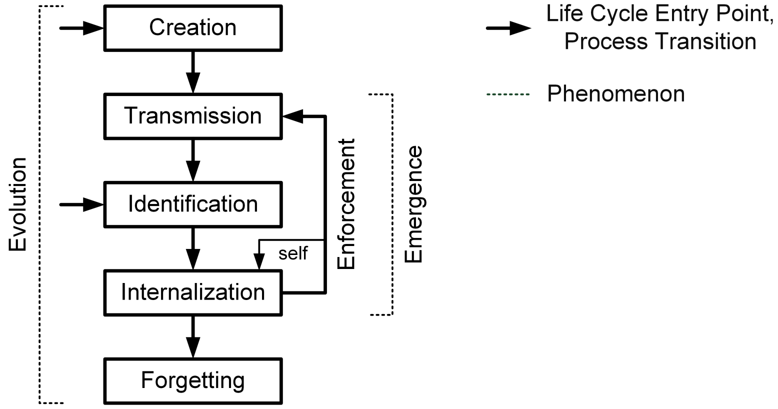

Figure 6: General Norm Life Cycle

As stated before, norms can either be explicitly created or identified at runtime (the corresponding right-facing arrows in Figure 6 mark these life cycles starting points). If created, norms are transmitted and identified.[12] As mentioned above, identification is not only initiated by transmission, but may involve the identification of an existing norm (e.g. by observation). Once internalized (by a complex internalization process that may contain elementary processes as laid out by Hollander and Wu [2011b]), norms can be reinforced, which may operate internally (e.g. based on motivational enforcement or elicited emotions),

[12] Note that we use terms synonymously for the ambiguous terminology in existing life cycle models as discussed before. In this case, the notion of 'identification' is identical to 'recognition'.

or be directed towards external targets. External enforcement requires the transmission of normative content, the subsequent identification and internalization by enforcement targets, and so on. This constitutes the norm's emergence. At any time, new norms can be created or identified, potentially causing change in the normative system by emerging and becoming salient. If cyclic reinforcements of a given norm cease, the norm loses its relevance and is incrementally forgotten. This second phenomenon can be understood as norm evolution. Both, emergence and evolution, are similar in that they represent phenomena (and could be construed as meta-processes in the epistemological sense[13]), but they vary in scope regarding the involved processes.

Norm Modification throughout Norm Life Cycle Hollander and Wu [2011b] discuss the modification of norms as part of the internalization process. However, we believe that the potential for norm modification, whether intentionally and systematic or not, arises during *any* form of transmission, internalization, or subsequent externalization (e.g. enforcement) of normative content.

This can involve the loss of information during transmission or simply transmission errors, leading to partial or simply wrong information. For example, ambient traffic noise may prevent bystanders from perceiving the scolding of jaywalkers or lead them to misconstrue the normative content (e.g. as a heated discussion).

Complementing potential modification sources during transmission, the identification of norms can be challenged by sensory biases that lead to a modified reproduction of normative content. Visual impairment, for example, may challenge or prevent an individual from capturing normative signals of relevance, such as the inability to observe a norm violation in the form of jaywalking.

During the internalization of norms, individuals can intentionally modify their interpretation of norms based on individual experience, background and aspirations. Hand-shaking, for example, can be interpreted as an acknowledgement of social status or objected to on the grounds of potential disease transmission. While the perceived action may be unambiguous (i.e. not manipulated during transmission and sensing), the individual may introduce an intentional bias, such as building a negative connotation with an internalized norm with an intent to change or abandon it.

This subjective perception of social reality extends to the unconscious realm, with an abundance of further mechanisms at work that drive individuals' biases in decision-making, belief formation and behaviour, as well as memory and social biases. Decision-making biases can be introduced by the oftentimes disproportionate perception of rewards and sanc-

[13] Our interpretation is in contrast to Hollander and Wu (Section 2.3) who use the term to describe end-to-end connections between elementary processes. They essentially consider regular processes and end-to-end processes as same natural kinds, and consequently do not allocate the operation of meta processes on a higher level of abstraction.

tions as well as an asymmetric risk tolerance (see e.g. Prospect Theory [Kahneman and Tversky, 1972]). An illustrative fact in line with this observation is that individuals are by magnitudes of thousands more likely to succumb to diseases from behavioural causes (e.g. lack of exercise, smoking) than terrorist attacks, yet fear the latter disproportionally more.

Further behavioural biases, for example, include paying selective attention to favourable information, as well as seeking for confirmation of conceptions and beliefs that we already hold (confirmation bias), such as the focus on information that 'validates' an existing norm. Memory biases are fundamentally concerned with humans' limited information processing capabilities (bounded rationality [Simon, 1955]), including limited information recall, the fading of memory over time, as well as our brain's ability to fill in of memory from imagination (false memories), all of which can lead to the distortion of internalized norms. Similarly, social effects can lead to biases with respect to the normative content, such as biases towards conformity with authority figures or ingroup members. Many of these systemic biases interact with human mechanisms for operating under uncertainty. Examples for such mechanisms include the use of stereotypes to ascribe characteristics to unknown individuals (implicit social cognition [Greenwald *et al.*, 2002]), or the application of irrational decision-making heuristics when acting under pressure ('gut feeling').

The presented selection of the cognitive biases is non-exhaustive, of course, but it offers a starting point for the exploration of cognitive influences that distort the interpretation of normative content during the norm internalization process.

Finally, norms can be modified based on the *characteristics of enforcement and enforcer*, generally affecting the salience and predictability of norms.

One fundamental determinant is the *valence* of enforcement, i.e. whether a norm is reinforced by rewards (such as a 'pat on the back') or punishments (such as scolding). As indicated earlier in the context of discussing cognitive biases, the nature of enforcement can modify norms. This includes the asymmetric impact of positive and negative sanctions (see e.g. Kahneman and Tversky [1972] and Baldwin [1971]), but also frequency, intensity and variation in enforcement. Infrequently reinforced norms are unlikely to gain high salience and may thus be easily foregone. Highly variable or inconsistent enforcement, however, interacts with individuals' risk affinity (e.g. promoting probabilistic norm compliance) but also involves the perceived level of fairness (e.g. inconsistent leadership behaviour in organisational environments [Sims and Brinkmann, 2003]), which can lead to the loss of norm commitment by norm subjects, or even active opposition.

Other influence factors on enforcement that can lead to norm modification include the *social relationship between enforcer and subjects*, but also the *nature of the enforcer*. As shown by Goette *et al.* [2006] and Horne [2007], increased social relationship (e.g. shared group membership) between enforcer and subjects correlates with the enforcement practice. However, the central or distributed nature of the enforcer can be decisive for the enforcement. Enforcers can be quasi-centralized and self-appointed (e.g. such as rules regarding dish washing procedures imposed by administrative secretary) and show predictable enforcement strategies (‘conventional sanctions’), whereas decentralized enforcement can be unpredictable with respect to the number of enforcers (e.g. unknown number of enforcers objecting to jaywalking), the applied strategies (e.g. gestures vs. scolding) and emerging dynamics (e.g. eruption into collective participation in humiliation), and thus lead to nuanced reinforcement and conceptualisation of the norm as more or less serious.

Complementing the misinterpretation of normative content based on sensory bias, enforcers can likewise cause a modification of normative content by sending ambiguous signals. Examples include the insufficient command of language to express a sanction appropriately or the confusion of terminology for reward and punishment (e.g. ‘awesome’ vs. ‘awful’).

Table 2 highlights the discussed potential causes for norm modification and associates those with individual processes. While this selection identifies potential modification sources, specific factors depend on the scenario, the capabilities of the transmission medium, as well as sensory and cognitive agent models and corresponding action capabilities. In addition to intentional modification, norms can thus essentially be modified whenever an individual interacts with its social environment, the effects of which can accumulate and drive the continuous evolution norms are subjected to, providing a starting point for exploring the emergence of divergent norms within separated social clusters.

Process	**Causes for Modification**
Transmission	Information Loss; Transmission Errors
Identification	Sensory Biases/Constraints
Internalization	Cognitive Biases; Intentional Modification
Enforcement	Choice of Enforcement; Characteristics of Enforcer(s); Relationship to Enforcement Target

Table 2: Potential Sources of Norm Modification

Summary In this section, we have proposed a general norm life cycle model that builds on the systematic comparison of existing life cycle models, harmonizes identified terminological and conceptual inconsistencies (see Section 2.6 for details), and introduces additional characteristics we deem relevant for a *general* norm life cycle model (e.g. norm identification as an alternative life cycle entry point).

While this proposed model highlights the essential processes of a general norm life cycle that we believe are necessary for its operationalization, it leaves the potential for the domain- or model-dependent refinement of individual processes, similar to Hollander and Wu 's model. However, this model integrates the commonalities of existing models, while offering a comprehensive and consistent reflection of norm dynamics found in the contemporary literature. It further provides a clear differentiation between processes and associated phenomena.

2.8 Discussion

Based on the condensed, yet comprehensive overview of selected existing normative life cycle models[14], we provided a systematic comparison and synthesized the identified essential components into a refined interpretation of the normative life cycle. However, the focus on individual processes of the life cycle models obscures two areas of development that *combine* individual processes to model norm dynamics comprehensively – the areas of *norm change* and *norm synthesis*. We will explore those specific areas in the following, before contextualising those with the proposed life cycle concept at the end of this article.

3 Norm Change

3.1 Overview

In the previous sections, we have seen different models that have been introduced in the literature to capture the life cycle of norms. These models consider the creation of norms, the processes that can facilitate their spreading, and the recognition (or learning) of norms by agents. Yet, we also know that in human societies norms can *change* over time. For example, on the occasion of the G8 summit in 2009 in Italy the Schengen treaty was suspended to guarantee the security of the local population and of the delegations, and then reinstated. In a similar way, normative systems in multi-agent systems must be able to evolve over time, for example due to actions of creating or removing norms in the system. However, the dynamic nature of norms in artificial systems is often not addressed in the simulation work on norms.

Norms are crucial in modeling agents' interactions. The definition of a normative multi-agent system that the community put forward at the first NorMAS workshop in 2005 is that “Normative MultiAgent Systems are multi-agent systems with normative systems in which agents can decide whether to follow the explicitly represented norms, and the normative systems specify how and in which extent the agents can modify the norms" [Boella *et al.*, 2006]. In order to ensure systems with autonomous agents, it is essential that norms can be

[14] Further life cycle models include the ones proposed by Andrighetto *et al.* [2013] and Singh [2014], but have been excluded from this comparison because of their highly abstract perspective or fine-grained computational focus.

violated (even though non-compliant agents are sanctioned). Because of the accent on the ability of the agents to modify norms, this definition was then known as “the normchange definition" of normative multi-agent systems.

The central problem of changing norms lead to two workshops on the dynamics of norms, the first one in 2007 in Luxembourg and the second one in 2010 in Amsterdam[15]. These two international workshops brought together researchers working on norm change from different perspectives. The revision of norms was also one of the ten open philosophical problems in deontic logic highlighted in Hansen *et al.* [2007] and further extended in Pigozzi and van der Torre [2017]. As we will see in the pages that follow, a consensus on a common framework to model norm change is still lacking.

3.2 From Law to Logic

Historically, the first approaches to norm change were driven by lawyers. For instance, at the 1981 international conference ‘Logica, Informatica, Diritto’ held in Florence (Italy), one of the conference sessions was explicitly dedicated to the problem of the abrogation of rules[16]:

> The abrogation of rules creates special problems in determining which is the ‘legal system in force’, as in the case of abrogation of the consequences of explicit rules and not of the rules themselves.

In the same years, a logic study of the changes of a legal code brought together three researchers coming from different backgrounds: Alchourrón, Gärdenfors and Makinson, respectively a legal theorist, a philosopher and a logician.

At the beginning, it was Alchourrón and Makinson who started investigating three types of change (Alchourrón and Makinson [1981; 1982]). The first type consists of the addition of a new norm (consistent with the other norms in the code) to an existing code. Such enlargement leads to the addition of the new norm to the code along with all the consequences that can be derived from it. The second type of change occurs again when a new norm is added, but now the new item is inconsistent with the ones already in the code. In this case we have an *amendment* of the code: in order to coherently add the new regulation, we need to reject those norms that conflict with the new one. Finally, the third change occurs when a norm is eliminated (technically, a *derogation*). In order for the elimination to be successful, however, also all other norms of the existing code that imply that norm have to be eliminated.

[15] http://www.cs.uu.nl/events/normchange2/

[16] When a norm is abrogated, its effects in the past still hold. This is different from the annulment of a norm, which also eliminates its effects in the past.

The approach of Alchourrón and Makinson was general: in the definition of change operators for a set of norms of some legal systems, the only assumption was that a norm is a formula in propositional logic. Thus, they suggested that "the same concepts and techniques may be taken up in other areas, wherever problems akin to inconsistency and derogation arise" ([Alchourrón and Makinson, 1981], p.147).

When Gärdenfors joined (at that time he was mainly working on counterfactuals), the trio became the founders of the well-known AGM theory, and started the fruitful research area of belief revision [Alchourrón *et al.*, 1985]. Belief revision is the formal study of how a theory (a deductively closed set of propositional formulas) may change in view of new information, which may cause an inconsistency with the existing beliefs.

Expansion, revision and contraction are the three belief change operations that Alchourrón, Gärdenfors and Makinson identified. *Expansion* is the addition of a new proposition that is not in conflict with the existing formulas in the theory. *Revision* is the addition of information that is inconsistent with the existing beliefs. In order to consistently add such information, all conflicting formulas have to be removed. Finally, *contraction* is the elimination of a formula from the theory.

The AGM theory provides a set of postulates for each type of theory change. There is an obvious correspondence between the three types of belief change and the three changes in a system of norms mentioned above. The link between theory change and change of a legal code was explicitly acknowledged by Alchourrón, Gärdenfors and Makinson:

> [...] theory *contraction*, where a proposition x which was earlier in a theory A, is rejected. When A is a code of norms, this process is known among legal theorists as the *derogation* of x from A. [...] Another kind of change is *revision*. [...] In normative contexts this kind of change is also known as *amendment*. ([Alchourrón *et al.*, 1985], p. 510)

It should be noted, however, that the AGM theory was mainly used for belief change. This is because beliefs and norms were both represented as formulas in propositional logic.

One of the first attempts to specify the AGM framework to tackle norm change was a paper by Maranhão [2001], presented at the 2001 ICAIL conference. The approach was inspired by Fermé and Hansson [1999]'s selective revision, where only part of the input information is accepted. Maranhão introduced a *refinement* operator, which refines an agent's belief set by accepting the new input under certain conditions. Refinement provides a tool to represent the introduction of exceptions to rules in order to avoid conflicts in normative systems (for instance in those cases where judges face new conditions which were not mentioned in the legal statute but turn out to be relevant in practical situations).

As we will see in the following pages, the belief revision approach has been recently reconsidered to represent and reason about norm change (see Section 3.4).

3.3 Semantic Approaches

Two main approaches to model norm change have been developed in the literature: semantic approaches inspired by the dynamic logic approach [van Ditmarsch and van der Hoek, 2007], and syntactic approaches where norm change is performed directly on the set of norms.

Among semantic approaches we find the dynamic context logic proposed by Aucher *et al.* [2009], which represents norm change (in particular the dynamics of constitutive norms[17]) as a form of model update. Starting from a modal logic of context [Grossi *et al.*, 2008], context expansion and context contraction operators are introduced. The intuition is that contexts can be seen as set of models of theories. Context expansion is thus linked to the promulgation of counts-as conditionals while context contraction is used for the abrogation of constitutive norms. Norms are statements of the kind "the fact α implies a violation". One of the advantages of this approach is that it can be used for the formal specification and verification of computational models of interactions based on norms.

A similar proposal is by Pucella and Weissman [2004], where operations for granting or revoking extensions are defined in a dynamic logic of permission. Aucher *et al.* [2009]'s framework is more general. Changes in the granting and revoking of permissions and obligations are more specific than the normative system change captured in Pucella and Weissman [2004]'s article.

3.4 Syntactic Approaches

3.4.1 Defeasible Logic

When new norms are created or old norms are retracted from a normative system, the changes have repercussions on obligations and permissions that such norms established. Obligations can change without removing or adding norms. For example, change in the world can lead to new obligations without changing the legal norms. For this reason, Governatori and Rotolo [2010] insist on the need to distinguish norms from obligations and permissions (as done in deontic logic).

Inspired by the legal practice, Governatori and Rotolo aim at a formal account of legal modifications. They use a syntactic approach, where norm change is an operation performed on the rules contained in the code. Such modifications can be implicit or explicit. Implicit modifications are the most common. They arise when new norms are introduced in the legal system and such norms conflict with existing ones. The new norms enforce a retraction of the old ones because, for example, have a higher ranking status, like a national law can

[17] Constitutive norms are rules that define an activity. For example, the institutions of marriage, money, and promising are systems of constitutive rules or conventions. As another example, a signature may count as a legal contract, and a legal contract may define a permission to use a resource and an obligation to pay.

derogate a regional law. Explicit modifications are obtained when norms that define how other existing norms have to be modified are added to the legal code.

In particular, the mechanisms of annulments and abrogations are studied. Annulment removes a norm from the code. It operates *ex tunc*: all effects (past and future) are cancelled. Abrogation too is a kind of norm removal but, unlike annulments, it applies *ex nunc*: it cannot operate retroactively, leaving their effects in the past hold.

The notion of abrogation is complex and there is no agreement among jurists on whether abrogations actually remove norms or not. In order to illustrate the difficulties, Governatori and Rotolo give the following example:

> If a norm n_1 is abrogated in 2007, its effects are no longer obtained after then. But, if a case should be decided in 2008 but the facts of the case are dated 2006, n_1, if applicable, will anyway produce its effects because the facts held in 2006, when n_1 was still in force (and abrogations are not retroactive). Accordingly, n_1 is still in the legal system, even though is no longer in force after 2007. ([Governatori and Rotolo, 2010], p. 159)

As seen in this example, the difficulty of abrogations comes from the fact that, in most cases, direct effects should be removed, but this is not necessarily the case for indirect effects. Clearly the temporal dimension is crucial in their formal representation, but it also makes the formalisation more cumbersome.

So Governatori and Rotolo first try to capture annulments and abrogations with theory revision in defeasible logic without temporal reasoning. Unfortunately, the result is not fully satisfactory as retroactivity cannot be captured. This is a crucial aspect as retroactivity allows to distinguish abrogation from annulment.

In the second part of the paper then, they use a temporal extension of defeasible logic to keep track of the changes in a normative system and to deal with retroactivity.

Norms have two temporal dimensions: the time of validity of a norm (when the norm enters in the normative system) and the time of effectiveness (when the norm can produce legal effects). As a consequence, multiple versions of a normative system are needed. In order to illustrate the problem, we recall this example from a hypothetical taxation law discussed in [Governatori and Rotolo, 2010]:

> If the taxable income of a person at January 31, for the previous year is in excess on $100,000$\$, then the top marginal rate computed at February 28 is 50% of the total taxable income. And this provision is in force from January 1. This rule can be written as follows:

$$(Threshold^{31Jan} \rightarrow HighMarginalRate^{28Feb})^{1Jan}$$

> Let us suppose that the last instalment for the salary was paid to an employee on January 4, and that it makes the total taxable income greater than the threshold stated above. We use $Threshold^{4Jan}$ to signal that the threshold of $100,000$\$ has been certified on January 4. [...] So let us ask what the top marginal rate for the employee is if she lodges a tax return on January 20. [...] [From] the point of view of January 20, the top marginal rate is 50%. Suppose now that there is a change in the legislation and that the above norm is changed on February 15, and the change is that the top marginal rate is 30%.
>
> $$(Threshold^{31Jan} \rightarrow MediumMarginalRate^{28Feb})^{15Feb}$$
>
> In this case if the employee lodges her tax return after February 15, the top marginal rate is 30% instead of 50%. ([Governatori and Rotolo, 2010], p. 173-174)

This example shows that what can be derived depends on which rules are valid at the time when we do the derivation, especially if rules can be changed. Thus, in order to keep track of the norm changes, Governatori and Rotolo represent different versions of a legal system.

3.4.2 Back to AGM

On May 19th, 1988 a three kilometres long bridge connecting the de Ré island in the Atlantic Ocean to France was inaugurated. Among the effects of such a convenient connection was that the price per square meters on the island flared up. Suddenly, farmers whose families had been living on the island sometimes since the XVth century, found they had to pay the wealth and large fortune tax, a tax directed to individuals who own assets of high net worth. Most of those farmers are retired people with low pension, living on the products on their fields of potatoes, asparagus and vines. In order to pay the wealth and large fortune tax, some had to sell part of their fields and endangered their retirements plans. This raised serious concerns on the unexpected implications of such tax and some people advocated a change of such law.

As we have seen, one of the motivations of the AGM theory of belief revision was the study of norm change. One may also argue that some of the AGM axioms (that have been criticized in the belief revision context) appear reasonable when applied to the legal discourse. The *success* postulate for revision, for example, imposes to always accept the new input. This postulate has been heavily criticized in the belief revision literature as irrational behaviours may result from it (consider, for example, an agent who receives a stream of contradicting inputs like $\phi, \neg\phi, \phi, \neg\phi, ...$). The success makes however sense in the legal context, when we wish to enforce a new norm.

As we have seen in the previous subsection, the explicit temporal representation and the use of meta-rules of Governatori and Rotolo [2010]'s approach resulted in complex logics. In order to reduce such complexity, Governatori *et al.* [2013] explored three AGM-like contraction operators to remove rules, add exceptions and revise rule priorities. Similarly to Governatori and Rotolo , this approach is rooted in the legal practice. The operators and the principles are illustrated with examples taken from the Italian Constitution and real decisions taken by the Italian Constitutional Court.

Boella *et al.* [2009] (subsequently extended in [Boella *et al.*, 2016b]) also reconsidered the original inspiration of the AGM theory of belief revision as framework to evaluate the dynamics of rule-based systems. Boella *et al.* [2016b] observe that if we wish to weaken a rule-based system from which we derive too much, we can use the theory of belief base dynamics [Hansson, 1993] to select a subset of the rules as the contraction of the rule-based system.

EXAMPLE 1.1 ([Boella *et al.*, 2016b], p.274) *Consider a rule-based system consisting of the following two rules:*

1. *If a then b*
2. *If b then c*

Assume we do not want to have c in context $\{a\}$*, whereas c can be derived by iteratively applying the first and the second rule. We can define rule base contraction operators that drop either the first or the second rule, or both.*

However, the next example illustrates that such rule contraction operators may not be sufficient.

EXAMPLE 1.2 ([Boella *et al.*, 2016b], p.274) *Assume d is an exception to c in context a. In that case, we may want to end up with a rule base consisting of the following two rules:*

1. *If* $a \wedge \neg d$ *then b, and*
2. *If b then c*

or a rule base consisting of the following two rules:

1. *If a then b, and*
2. *If* $b \wedge \neg d$ *then c.*

In other words, in some applications, we may need to change *some of the rules. In particular, rule contraction may assume a* rule logic *which informs us that the rule 'if a then b' implies the rule 'if $a \wedge \neg d$ then b', or that 'if b then c' implies the rule 'if $b \wedge \neg d$ then c'.*

Thus, even if base contraction is the most straightforward and safe way to perform a contraction, it always results in a subset of the original base, which sometimes means removing too much. Take, for example $\{(a,x)\} \div (a,x) = \{\}$, where $\div$ denotes the contraction operator. Thus, under base contraction, the only result is to throw away the rule. But under AGM one can put a weaker rule. For instance, if (a,x) is the rule "If an individual owns land for more than 1.3 million Euros (a), then he must pay the wealth and large fortune tax (x)". To avoid problems as those on de Ré island, we may wish to change the law by introducing an exception, like $\{(a,x)\} \div (a,x) = \{(a \wedge b, x)\}$, where b stays for people with high income.

This was one of the motivations of Boella *et al.* [2016b]. In their abstract approach, rules are pairs (a,x) of propositional formulas and a normative system R is a set of pairs. Several logics for rules are considered by resorting to the input/output logic framework developed by Makinson and van der Torre [2000; 2003].[18]

Rules allow to derive formulas, that is, obligations and prohibitions in a normative system. The factual situation (called *context* or *input*) determines which obligations and prohibitions can be derived in a normative system. Formally, in the input/output notation: if $(a,x) \in R$ then $x \in out(R,a)$. This means that, according to the normative system R, in context a, the formula x is obligatory. The idea is that a is the input (or context) and x is the output. Of the operations defined semantically and characterized by derivation rules in Makinson and van der Torre [2000], three operations are considered in Boella *et al.* [2009; 2016b]: simple-minded, basic, and simple-minded reusable.

In order to generalize the AGM postulates for normative change, a rule set is taken to be a set of rules closed under an input/output logic. Rule expansion, rule contraction and rule revision in the input/output framework are then defined. Similarly as for the belief change case, the definition of rule expansion is unproblematic. Here, the legislator wishes to add a new rule that does not conflict with the existing ones. Rule contraction and rule revision, on the other hand, are more interesting.

AGM postulates for expansion, contraction and revision are reformulated for rule expansion, rule contraction and rule revision. It turns out that (surprisingly) the postulates for rule contraction are consistent only for some input/output logics, but not for others. On the positive side, the proof theory of rule change was shown to be closely related to the proof theory of permissions from an input/output perspective [Boella *et al.*, 2016b].

[18] Maranhão [2017] employs input/output logics and belief revision principles to model legal interpretation. Judicial doctrine is seen as theory change, where rules and values need to be revised to obtain a coherent system.

The translation from the AGM contraction postulates to the postulates for rule revision turned out to be more difficult. One of the difficulties was the definition of the negated input (roughly corresponding to $\neg(a,x)$) and the inconsistent set of rules in input/output logic (which would correspond to an 'incoherent' system of rules in the normative systems paradigm).

Postulates for (belief and rule) revision and (belief and rule) contraction are independent. No contraction operator appears in the revision postulates, and no revision operator appears in the postulates for contraction. Yet, the Levi identity and the Harper identity defined respectively the belief revision operator as a sequence of contraction and expansion, and the belief contraction is defined in terms of belief revision.

Using the Levi identity, rule revision was defined in terms of rule contraction. The operators so defined were shown to satisfy the AGM postulates. For the Harper identity, however, the question is still open [Boella *et al.*, 2016b].

A similar approach to Boella *et al.* [2009; 2016b]'s has been proposed by Stolpe [2010]. There, AGM contractions and revision are used to define derogation and amendment of norms. In particular, the derogation operation is an AGM partial meet contraction obtained by defining a selection function for a set of norms in input/output logic. Norm revision defined via the Levi identity characterize the amendment of norms. Stolpe can thus show that derogation and amendment operators are in one-to-one correspondence with the Harper and Levi identities as inverse bijective maps.

3.5 Computational Mechanisms of Norm Change

Beside the theoretical investigations to norm change presented in the previous sections, few work exist on the computational mechanisms of norm change.

The drawback of determining norms at design time is that unforeseen situations may occur and the system cannot adapt to the new circumstances. The approach proposed by Tinnemeier *et al.* [2010] tackles this problem by allowing the modification of norms at runtime, so that a programmer can stipulate when and how norms can be modified. In Tinnemeier *et al.* [2010]'s framework norms can be modified by external agents as well as the normative framework.

The proposed norm change mechanism is system-dependent and enforcement-independent. The first principle states that who can change norms, how and when norms may be changed depends on the system. The authors justify this first principle by recalling the clause that a normative system must "specify how and in which extent the agents can modify the norms", as in the definition proposed at the first NorMAS workshop in 2005. The second principle ensures that the norm change and the norm enforcement mechanisms should be defined independently. This is to increase the readability and manageability of the program.

Two types of norm change rules are defined. The first type is used to change instances of norms without modifying the norm scheme. These rules define the circumstances under which some norm instances have to be removed to be replaced by other norm instances. The second type of rules is used to alter norm schemes. As for the first type, these rules define under which circumstances norm schemes are to be changed by retracting some norm schemes and adding others.

What happens to the instances already instantiated, when the underlying norm scheme is changed? Tinnemeier *et al.* [2010] observe that there are situations in which we want to leave the instantiated instances unchanged, and others in which it makes sense to apply the change retroactively. Thus, two types of norm scheme change rules are given. Finally, building on [Tinnemeier *et al.*, 2009], the syntax and operational semantics of the programming language are given.

Previous work on norm change at runtime includes [Bou *et al.*, 2007; Campos *et al.*, 2009]. Bou *et al.* [2007] also consider the problem of adapting a system to novel and unpredictable circumstances. To this end, they present an approach to enable normative frameworks (called “electronic institutions" in [Bou *et al.*, 2007; Campos *et al.*, 2009]) to adapt norms to agents’ behaviour changes as well as to comply with institutional goals. The norm change mechanisms of Bou *et al.* [2007] allow to modify existing norms. Unlike [Tinnemeier *et al.*, 2009], new norms cannot be introduced nor can existing norms be removed. Another difference is that Bou *et al.* [2007] use a quantitative approach to represent the environment and the agents.

Campos *et al.* [2009] approached the difficulty of how to adapt a normative system to the changes of its agents’ behaviour by adding situatedness and adaptation (two properties usually characterising agents) to the system. The result is a system that can make changes and that can also adapt to changes. As in Bou *et al.* [2007]’s approach, the aim is to modify agent coordination to enhance the system’s performance in attaining institutional goals.

Even though Boella and van der Torre [2004]’s approach is theoretical, it shares some similarities to the works presented here. Starting from the distinction between regulative norms (that indicate what is obligatory or permitted) and constitutive (or count-as) rules (that define an activity), they use constitutive rules to create new norms as well as to define what changes the agents can introduce. As in the norm instance change rules and norm scheme change rules of Tinnemeier *et al.* [2010], constitutive and regulative rules in Boella and van der Torre [2004] are modelled as conditional rules specifying when a norm can be changed and what the consequences are.

3.6 Discussion

In this short excursus we have seen that the first formal investigations of changes in a legal code had roots in logic, namely in the AGM framework. This line of research has been reconsidered, notably in the works of Governatori and Rotolo [2010; 2013], Stolpe [2010], and Boella *et al.* [2009; 2016b], often coupled with non-classical logics such as defeasible logic or input/output. Another direction has been to follow a semantic approach inspired by dynamic logic, as in Pucella and Weissman [2004] and Aucher *et al.* [2009]. Finally, besides the theoretical investigations, some work on the computational mechanisms of norm change has been done, like Tinnemeier *et al.* [2010], Bou *et al.* [2007] and Campos *et al.* [2009].

Norm change is a fairly recent research theme in the NorMAS community. The first international workshop explicitly dedicated to the dynamics of norms was held in 2007. This observation can in part explain the lack of consensus around a common theoretical framework. But it probably does not explain it completely. Other reasons may reside in the limits of abstract frameworks like AGM, even when combined with with richer rule-based logical systems, in the difficulty to capture and distinguish norm change from changes in obligations, and again in the elusive character of legal changes in the real world. Recent developments in legal informatics may help casting light on norm dynamics. Works on legal document and knowledge management systems (like the EUNOMOS project [Boella *et al.*, 2016a]) allow, for example, to keep track of (implicit and explicit) changes in the legislation. Although these works provide some first steps in the understanding of the dynamics of normative systems, much still remains unexplored.

4 Norm Synthesis

The second theme of norm synthesis has a long-standing history but has experienced a recent revival of attention. While norm change primarily focuses on the logical implications of the modification of existing (legal) norms over time, norm synthesis puts a stronger emphasis on how (social) norms emerge and converge in the first place, and how they can be identified.

4.1 Foundations

Norm synthesis is inspired by the area of program synthesis (i.e. generating a program from a given specification [Manna and Waldinger, 1980]), but, in contrast to the former, shifts the focus to the coordination of autonomously operating agents. The specific purpose of norm synthesis is thus to identify an optimal set of norms (a normative system) to coordinate individuals' behaviours in a multi-agent system. The 'optimality' of a solution depends on the specified objectives, such as the minimal set of norms to facilitate coordination [Fitoussi and Tennenholtz, 2000].

Shoham and Tennenholtz [1992b; 1995]'s work on synthesis of social laws is considered the initial work in the area of norm synthesis. They propose a general formal model to identify a set of social laws at design time (offline) to assure the coordinated operation of structurally uniform agents. They showcase this approach by 'handcrafting' a set of social laws that guarantee collision-free coordination in a grid-based traffic scenario ('rules of the road'[19]), instead of determining action prescriptions for each possible system state. However, they also show that the automated synthesis for offline approaches is NP-hard [Shoham and Tennenholtz, 1995], challenging the generalizable application. Onn and Tennenholtz [1997] propose a general solution for the synthesis problem for scenarios that can be represented as biconnected graphs by reducing synthesis to a graph routing problem. Fitoussi and Tennenholtz [2000] further introduce qualitative characteristics for synthesized social laws, such as their *Minimality* and *Simplicity*. As alluded to before, minimal social laws seek to specify fewest possible restrictions on agents' behaviours, thus giving individuals the greatest possible autonomy, while maintaining coordination in the overall system. An extremely restrictive social law would effectively prescribe any action an agent could take in any given situation (e.g. to walk on the right side of a footpath in a given direction, or even more restrictive, prescribing specific navigation routes between different locations), thus removing any form of autonomy on the part of the agent. A minimal social law (e.g. not to step on the road), in contrast, would retain the agent's ability to pursue its own goals, as long as it is compatible with the system objectives (e.g. avoiding collisions between cars and pedestrians). In a more recent approach, Christelis and Rovatsos [2009]'s work on automated offline norm synthesis addresses the complexity problem by identifying prohibitive states in incomplete state specifications that are generalized across the entire state space. It is important to note that these early approaches to norm synthesis do not consider or tolerate any form of violation; unlike most subsequent work, their conceptions of social laws describe hard constraints agents cannot forego.

The shift towards refined norm interpretations that emphasizes the interactionist over legal perspective (and thus regulation over regimentation) [Boella *et al.*, 2008] has stimulated a differentiated treatment of rewards and sanctions as mechanisms of social enforcement. This sociologically-inspired norm perspective drove the exploration of associated influence factors (such as memory and connectivity), along with a movement from *offline* to *online* norm synthesis, resulting in two subfields. *Convention/Norm Emergence* (which we will differentiate later) emphasize mechanisms that influence the convergence on norms or conventions, whereas work we cluster under the label *Identification* concentrates on the mechanics of detecting and synthesising norms in the first place. The latter can further be subdivided into approaches that rely on a centralized or decentralized operation, that is, approaches that use a central entity to synthesize norms, or delegate the generalization and integration of identified norms to the agents themselves. Figure 7 provides a schematic overview of the outlined structure of the research field. Overall, the subfields of norm synthesis cover the

[19] This *de facto* reference scenario has been adopted and refined in large parts of subsequent work on norm synthesis.

notion of norms in the broad sense (i.e. as institutions), ranging from self-enforcing conventions via socially enforced norms to centrally enforced social laws or rules. In the following, we will discuss selected contributions to the area of norm synthesis, with a particular focus on approaches that emphasize the detection and identification of norms.

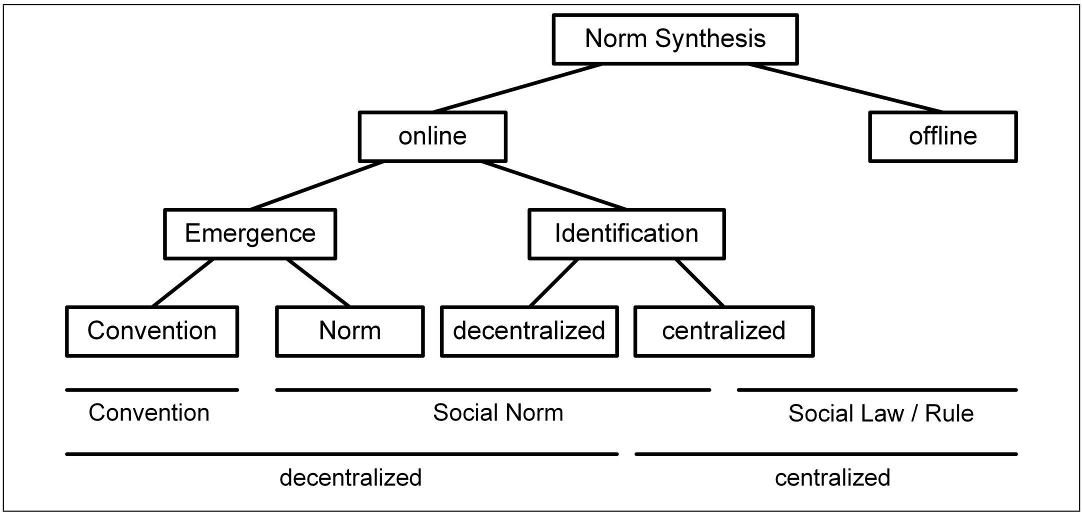

Figure 7: Taxonomy of Norm Synthesis Approaches

4.2 Synthesis as Norm/Convention Emergence

Research efforts in the area of *norm emergence* put particular concentration on an understanding of the contextual conditions and mechanisms that bring norms about, including their distributed nature. Instead of relying on a centralized entity to determine norms a priori or embedding hard-coded (offline designed) norms into individuals, norm emergence affords the decentralized collaboration of agents to converge on commonly accepted social norms.

Explored mechanisms include:

- Memory size (e.g. Villatoro *et al.* [2009])
- Network topologies and dynamics of relationships (e.g. Savarimuthu *et al.* [2009], Villatoro *et al.* [2009], Sen and Sen [2010], Sugawara [2011], Villatoro *et al.* [2013])
- Clusters (e.g. Pujol *et al.* [2005])
- Interaction-based social learning (e.g. Sen and Airiau [2007], Mukherjee *et al.* [2007; 2008], Airiau *et al.* [2014])

- Lying (e.g. Savarimuthu *et al.* [2011])
- Dynamic sanctions (e.g. Mahmoud *et al.* [2012; 2015])
- Hierarchical structures with varying levels of influence (e.g. Franks *et al.* [2013; 2014], Yu *et al.* [2013; 2015])

Further contributions in the area of norm emergence include algorithms for distributed decision-making to arrive at a shared lexicon [Salazar *et al.*, 2010] or shared sets of tags [Griffiths and Luck, 2010].

The decentralized operation of norm emergence places an emphasis on larger number of agents and their direct interaction in favour of cognitive capability and central coordination. Consequently, the computational complexity of individual agents is limited and the applied norm representations are mostly abstract in the form of converging strategy choices in coordination games or string-based representations; the normative content is symbolic and can only be inferred from the motivating scenario. In addition to the abstract normative content, in most cases, agents converge on a single norm (with exception of Savarimuthu *et al.* [2009] and Sen and Sen [2010]). In addition, most scenarios sustain the emerging norm without explicit enforcement, thus representing *self-enforcing conventions* as opposed to *externally enforced social norms*, affording the differentiation into *Convention Emergence* and *Norm Emergence*.

Following the exploration of the emergence strand of norm synthesis, we will turn to the identification strand that captures norm synthesis processes in a narrow sense, primarily focusing on the detection, identification, and integration of norms into consistent normative systems.

4.3 Synthesis as Identification

Work that identifies and synthesizes norms at runtime can be differentiated into centralized approaches, which interpret norm synthesis in the original spirit of identifying centrally managed system-wide norms, and decentralized ones that analyze the inception of norms from a bottom-up perspective.

A series of centralized online norm synthesis approaches that follow the tradition of Shoham and Tennenholtz has been spearheaded by Morales *et al.* . In their work, Morales *et al.* [2013] propose the *Intelligent Robust Norm Synthesis* mechanism dubbed IRON in an adapted version of the grid-based 'rules of the road' scenario originally introduced by Shoham and Tennenholtz that focuses on coordination in traffic junctions. Agents have a limited observational range and move in travel direction, unless constrained by imposed norms. IRON continually monitors traffic participants' behaviour. When detecting collisions, IRON identifies the underlying conditions (e.g. car approaching from the right) and introduces a norm that prevents a similar event from reoccurring (e.g. by introducing an obligation to stop whenever facing a car to one's right). These centrally generated and managed norms (which make those effectively rules or social laws) are imposed upon all traffic participants, thus progressively moving towards a stable collision-free normative system.

To prevent overregulation from introducing too many specific norms based on individual observations, IRON attempts to *generalize* norms based on their shared preconditions by selectively ignoring a specific norm's partial precondition. The generalized norm is evaluated at runtime by detecting eventual recurring collisions, in which case the original specific norms are deemed relevant and are reinstated. To determine the *effectiveness* of given norms, IRON further monitors their activation, and ascribes frequently applied norms higher effectiveness. To identify *necessary* norms, Morales *et al.* [2013] (unlike Shoham and Tennenholtz [1995]'s social law approach) make use of the agents' ability to violate norms, which enables IRON to identify imposed norms that are actually necessary to maintain coordination and remove unnecessary ones (i.e. norms whose violation does not produce collisions).

Morales *et al.* [2014] successively introduce further iterations of their approach (dubbed SIMON) that consider structural diversity of norm participants (e.g. by introducing emergency vehicles) and refined mechanisms for norm generalization with specific focus on minimizing the necessary simulation runtime to produce a collision-free normative system Morales *et al.* [2014; 2015c] . Their following system iteration, LION [Morales *et al.*, 2015b], includes the focus on the identification of semantic relationships between norms, so as to produce fewer, more general norms (liberal norms) that maximize the norm participants' autonomy.

This series of works on norm synthesis highlights the advantages of centralized approaches not only to identify norms, but to integrate those. In this interpretation, synthesis involves an explicit analytical effort to integrate individual norms into a coherent normative system, producing semantically meaningful complex coordination outcomes, beyond a coordinated strategy choice as observed in most norm emergence approaches. Consequently, a comprehensive approach to norm synthesis captures life cycle processes that include identification, as well as internalization and forgetting of norms, thus covering processes that are associated with the evolution of norms over time (see Section 2.6). Processes such as spreading and enforcement, characteristically associated with the work on norm emergence, are secondary.

Riveret *et al.* [2014]'s transfiguration approach takes an incremental step towards decentralized systems by endowing individual agents with learning capabilities enabling them to infer behavioural prescriptions from stochastic games. Being grounded in the field of computational justice, their approach marries bottom-up dynamics (transfiguration of experience into prescriptions) with notions of self-governance by means of collective action (voting). The voting process is initiated once all agents have submitted their inferred (and preferred) prescriptions, the most common of which is put forth as a motion. Agents are then invited to vote based on the perceived purposefulness of the prescription content, which is abstractly represented using a notion of global and individually perceived *potential*. Since the purpose of the voting process (in the spirit of self-governance) is to promote globally useful prescriptions, the agents decide probabilistically based on the alignment of the candidate prescription's individual and global potential. Once adopted, the prescription becomes a self-imposed rule of that society.

This work emphasizes the computational representation of social processes that enable self-governance by retaining high levels of decisional autonomy on the part of the society members, while abstractly providing centralized decision-making and enforcement inspired by real societies. Beyond the conceptual integration of bottom-up and top-down governance processes, this contribution emphasizes the efficiency benefits associated with centrally coordinated collective decision-making.

Contributions that shift the perspective away from approaches that emphasize effective coordination towards individual-centric operations can be captured under the umbrella of *decentralized online norm synthesis*. In addition to the focus on the individual as an entity of concern, in principle these approaches lend themselves well for explorative scenarios with a broader (if not open) range of actions than used in the centralized coordination scenarios. Research efforts related to this cluster include Andrighetto *et al.* [2007; 2010] as well as Savarimuthu *et al.* [2010b; 2013a]. We will not discuss these works in great detail at this stage as we covered those in the context of norm creation in Hollander and Wu [2011b]'s life cycle model (see Section 2.3). Instead, we will concentrate on contributions that treat norm synthesis as a holistic process involving multiple life cycle processes.

An important work in this area is Savarimuthu *et al.* [2013b]'s work on norm recommendation. Their approach is motivated by the identification and recommendation of an existing system's norms to newcomers, which can operate in a centralized or decentralized fashion. Their system combines norm identification, classification and life cycle stage detection in order to recommend the existence and relevance of observed norms. The initial step of norm detection operates on a continuous stream of events by identifying recurring event episodes that are terminated with a sanction signal. The algorithm collects event episodes of varying window sizes in order to establish the subset of actions that provoke a sanction signal and identifies those as candidate norms. In the second step, norms are classified with respect to their salience. For this purpose, the mechanism tracks both the invocation of actions contained in the candidate norms as well as the frequency of punishments as a response to action activation. By ranking these measures, the mechanism classifies norms by salience, where the existence of punishment is indicative of higher levels of salience, as opposed to mere action activation. A further step emphasizes the long-term perspective and attempts to identify a norm's life cycle stage (*life stage*), with possible stages being emerging, growing, maturing, declining, and decaying. The system monitors norms' punishment probabilities over time and evaluates those with respect to given successive thresholds associated with emergence (frequency of activation) and growth, based on which it infers the life stage. For example, norms that have experienced an increase in punishment between two time intervals but remain between the emergence and growth thresholds, are considered growing. The system uses heuristics that use the established measures for salience and life stage as an input to *recommend* the existence of a given norm.

Similar to Morales *et al.* [2015a]'s works, Savarimuthu *et al.* [2013b]'s synthesis approach allows the identification of multiple norms, along with a quantitative measure of salience that is comparable with Morales *et al.* [2013]'s notion of effectiveness and necessity. Savarimuthu *et al.* [2013b]'s approach further includes a systematic classification of norms with respect to their life cycle stage, thus emphasising the long-term perspective. However, unlike Morales *et al.* [2015a], this work relies on an abstract string-based norm representation and does not consider semantic relationships between norms, thus preventing operations such as generalization or substitution of norms.

The final approach we present under the umbrella of norm synthesis takes an intermediate stance by operating decentralized while maintaining meaningful norm representations. Frantz *et al.* [2014c; 2015] propose a norm generalization approach that operates on individual observations. At its core, this approach is motivated by individuals' tendency to subconsciously develop stereotypes as decision-making shortcuts they can use when encountering unknown interaction partners. To facilitate this generalization, the mechanism relies on uniform structural representations of actors, actions and norms based on Nested ADICO (nADICO) [Frantz *et al.*, 2013; Frantz *et al.*, 2015], a rule-based norm representation that builds on the *Grammar of Institutions* [Crawford and Ostrom, 1995] and affords the explicit representation of structural institutional regress [Frantz, 2015], i.e. the nested interdependency of sanctions and corresponding metanorms. As a first step, observations are aggregated based on shared observable attributes as well as subsets thereof (higher generalization levels), forming the basis to synthesize descriptive norms (or conventions) the observer attributes to observed groups of individuals. To infer injunctive norms from observations, individuals further track corresponding reactions to ascribe the generalized action sequences normative character and interpret the generalized reactions as social consequences (i.e. rewards or sanctions). The frequency and intensity of observations indicate a norm's salience by mapping it onto a continuous deontics conception (*Dynamic Deontics* [Frantz *et al.*, 2014a]) that spans from prohibition via permission to obligation, the *deontic range* of which is unique for each agent and determined by its previous experience. In addition to the extremal cases, this concept introduces intermediate stages along this continuum (e.g. obligations that are omissible and can be exceptionally foregone), a principle that is used to reflect the subjectively perceived priority of a given norm, and implicitly solves potential norm conflicts.

In contrast to the approach by Morales *et al.* [2015a], this work does not solve a specific coordination problem, but introduces a fully decentralized approach to understand agents' behaviours by inspecting their *subjective understanding* of a scenario's normative content, thus shifting it into closer proximity to emergence-based approaches. Similar to Morales *et al.* [2015a] (but unlike Savarimuthu *et al.* [2013b]), this approach uses a comprehensive human-readable norm representation (as *institutional statements*) and allows the identification of norm relationships by generalizing individual observations. The uniform norm representation further permits the analysis on arbitrary social aggregation levels (e.g. group, society).

Table 3 provides a chronological overview of all identified norm synthesis approaches based on the characteristics introduced at the beginning of this subsection (see Figure 7), including the nature of norm (convention, norm, rule, social law), central coordination and the ability to produce or identify multiple norms.

Contribution	**Institution Type**	**Centralized**	**Offline**	**Single Norm**
Shoham and Tennenholtz [1995]	Social Law	yes	yes	no
Pujol *et al.* [2005]	Convention	no	no	yes
Sen and Airiau [2007]	Convention	no	no	yes
Savarimuthu *et al.* [2007; 2008a]	Norm	no	no	no
Mukherjee *et al.* [2007; 2008]	Convention	no	no	yes
Christelis and Rovatsos [2009]	Social Law	yes	yes	no
Villatoro *et al.* [2009]	Convention	no	no	yes
Urbano *et al.* [2009]	Convention	no	no	yes
Sen and Sen [2010]	Convention	no	no	yes
Griffiths and Luck [2010]	Norm	no	no	no
Sugawara [2011]	Convention	no	no	no
Mahmoud *et al.* [2012]	Norm	no	no	yes
Morales *et al.* [2013]	Social Law	yes	no	no
Franks *et al.* [2013]	Convention	no	no	yes
Villatoro *et al.* [2013]	Convention	no	no	yes
Savarimuthu *et al.* [2013b]	Norm	both	no	no
Mihaylov *et al.* [2014]	Convention	no	no	yes
Airiau *et al.* [2014]	Convention	no	no	yes
Morales *et al.* [2014]	Social Law	yes	no	no
Riveret *et al.* [2014]	Norm / Rule	yes	no	no
Frantz *et al.* [2014c; 2015]	Norm	no	no	no
Morales *et al.* [2015b]	Social Law	yes	no	no
Mahmoud *et al.* [2015]	Norm	no	no	yes

Table 3: Overview of Norm Synthesis Approaches

4.4 Contextualization with the General Norm Life Cycle Model

At the current stage, norm synthesis presents itself as a diverse field that is driven by varying objectives. Apart from the historical separation into offline and online approaches, we can identify a cluster of existing approaches that either concentrate on the:

- Investigation of factors and circumstances that promote norm adoption (emphasizing macro-level outcomes), or
- Mechanisms for the runtime identification, generalization, implementation, and integration with established norms (emphasizing micro-level mechanisms).

Relating these approaches to individual life cycle processes of the general norm life cycle model (see Section 2.7) as shown in Figure 8, we can observe that emergence-based approaches emphasize spreading/transmission mechanisms (e.g. type and dynamic nature of network topologies, hierarchical structures, social learning, memory size) along enforcement characteristics (e.g. sanctioning, lying).

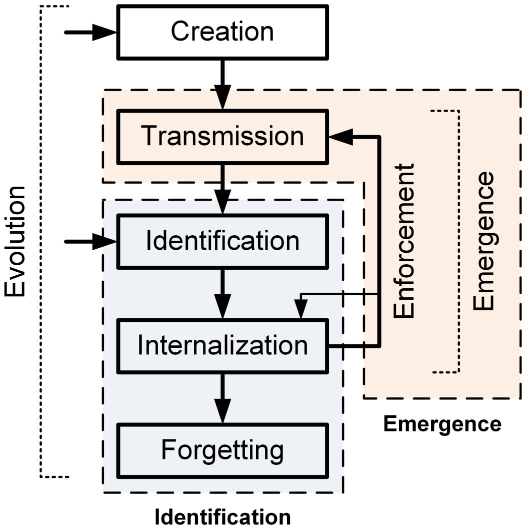

Figure 8: Norm Synthesis Approaches and Related Life Cycle Processes

The second group of mechanisms emphasize the detection and identification of existing norms. Deductive tasks for the generalization of comprehensive normative systems are related to a complex norm internalization process (such as the one conceptualized by Hollander and Wu [2011b]), since it represents a composite process that merges newly discovered norms and existing sets of norms, which requires the ability to modify, generalize and integrate norms. The synthesis of normative systems further relies on the ability to discard or forget norms.

Despite the comprehensive coverage of different life cycle stages, the review of existing approaches indicates gaps. An important central topic that has found limited explicit attention in current approaches is the detection of norm conflicts, an aspect with a strong relation to the norm internalization process. Riveret *et al.* [2014], as well as Savarimuthu *et al.* [2013b], treat norms independently without considering their relationship to existing norms. Frantz *et al.* [2015] and Morales *et al.* [2015a] include generalization processes and mechanisms to accommodate conflicting or competing norms, but only Morales *et al.* [2015b] perform explicit detection of norm relationships such as complementarity and substitutability. An area that has found recent attention is the focus on *dynamic normative systems* [Huang *et al.*, 2016] in which the normative environment itself is not considered static, but changes over time, and thus requires agents to revise their normative understanding in order to accommodate those changes. Initial work by Huang *et al.* [2016] analyzes the associated complexity of norm recognition and synthesis.

4.5 Discussion of Challenges and Future Directions

In this section, we provided a comprehensive discussion of the historical roots of norm synthesis, the shifts from offline to online synthesis, and the subsequent differentiation into more implicit emergence-focused and more explicit identification-centric approaches. We further discussed a set of relevant contributions to the latter identification strand of norm synthesis. However, apart from surveying the field, this comprehensive overview of the area of norm synthesis allows us to identify areas which we believe deserve further attention.

Reviewing the strands of (online) norm synthesis, an outstanding development is the systematic integration of both strands by enriching emergence-based approaches with richer micro-level architectures that incorporate components of identification-based mechanisms. For identification-based approaches, this implies a stronger focus on generalizable representations of norms and social structures beyond specific scenarios. The marriage of both strands provides a basis for more realistic representations of social scenarios, with *emergence* sponsoring the insight on how to structure interaction in social environments, and *identification* providing mechanisms to develop complex, yet consistent normative systems as we encounter them in the real world.

We further believe that the exploration of dynamic normative systems represents an important research direction if we aim towards the use of norm synthesis in real-world applications (e.g. robotics). It further has the potential to link the theoretical contributions developed in the area of norm change, e.g. modelling changes in legal systems (as discussed in Section 3), with the mechanisms that facilitate the identification, generalization and integration of corresponding operational norms developed in the area of norm synthesis.

Looking beyond the scope of contemporary work, an important challenge for the successful adoption of norm synthesis is the identification and development of application domains that enable the use of these techniques in realistic scenarios, both involving the extent and complexity of available data. In this context, a challenge that all contemporary approaches to norm synthesis share is their operation on structured data. Making unstructured, noisy or semi-structured data (such as found in big data) accessible under consideration of the complexity limitations of current norm synthesis approaches will increase its relevance for real-world applications. Specific examples include the automated the extraction of norms from large and diverse real-world data corpi, as well as performing online norm synthesis, e.g. for the ad hoc inference of normative understanding in the context of robotics or digital assistants.

5 Summary, Conclusions and Outlook

In this article, we have provided an overview of the contemporary perspective on norm dynamics, with a particular focus on norm change and norm synthesis as important active research fields in multi-agent systems.

The research around norm change (Section 3) has resulted in a comprehensive exploration of logical challenges associated with the representation of changing social and legal norms, such as temporal implications of changing laws and an adequate formal translation of the notion of an incoherent normative system. At this stage, the relatively young but promising field has yet to find a shared consensus on the theoretical foundation to provide the platform for the systematic application of its contributions in the context of normative multi-agent systems as well as other disciplines.

Research in the area of norm synthesis (Section 4) concentrates on the analysis of factors that contribute to emerging norms (norm emergence) as well mechanisms to detect existing norms (norm identification). This field has experienced a revival with the recent focus on the synthesis of normative systems at runtime (online) – as opposed to the traditional offline approach. In addition, the field features an increasing number of approaches that favour decentralized over centralized approaches or combine both approaches and use social choice mechanisms for the integration of bottom-up and top-down perspectives on norm synthesis.

To understand the developments in both fields, we initially presented an overview of approaches that define the norm life cycles (Section 2), while providing an overview of the contemporary state of current contributions associated with individual life cycle processes. We further systematically compared the surveyed life cycles based on involved processes and norm characteristics, while identifying abstract phases of the norm life cycle. From this analysis, we extracted the essential processes and integrated those in a *general norm life cycle model* that reflects the contemporary view on norm emergence. The refined model resolves terminological and conceptual inconsistencies/omissions identified in the existing life cycle models. It further suggests that external influence factors can lead to norm modification throughout all stages of the norm life cycle, and, unlike earlier models, distinguishes between normative processes and associated phenomena.

Since this article specifically concentrates on the *modeling of norm dynamics*, we do not capture the wider technical and philosophical implications of norm dynamics, such as the dealing with normative conflicts and violations (see article 'Modeling Normative Conflicts in Multi-Agent Systems' in this volume), aspects of norm autonomy (see Verhagen [2000]), and the role of trust for the functioning of norms (see Andrighetto *et al.* [2013]).

Surveying individual contributions to the field of NorMAS in general – and to the areas of norm change and synthesis in particular – we can observe a tendency to apply richer norm conceptions that span across multiple norm life cycle processes. As a result, developed systems produce increasingly dynamic behaviour. This includes a) the identification of norms at runtime, b) the change of norms over time, and c) their potential decay and substitution.

These observations highlight an important progression for the wider discipline, since it positions the current development on the roadmap laid out in the 2007 Dagstuhl NorMAS workshop that identified five levels in the development of normative multi-agent systems (see Boella *et al.* [2008]):

- Level 1 – Off-line designed norms
- Level 2 – Explicit norm representations that can be used for communication and negotiation
- Level 3 – Runtime addition, removal and modification of norms
- Level 4 – Embeddedness in social reality
- Level 5 – Development of moral reality

The first three levels are undisputed – the shift towards dynamic creation (Level 3) is reflected in numerous contributions to the field. However, the ability of agents to identify and synthesize norms in their social environment at runtime, the ability to engage in social choice processes, as well as agents' compliance in dynamic normative systems provide the basis to make agents active participants in shaping social reality, and thus moves them closer to the fourth development level (without discussing the associated challenges at this stage – for details see Boella *et al.* [2008]).

Fundamentally, this integration of normative concepts in social reality cannot be dissociated from the consideration of ethical and moral concerns as suggested for the last level – the development of moral reality by assuming moral agency. This resonates with contemporary developments, such as the productive use of autonomous cars, increasing automation of the workforce via robotics, decentralisation of autonomy (e.g. in distributed ledger technology), along with the revived societal debates around the impact of artificial intelligence (e.g. recall the debates around universal base income). This necessity to address the embeddedness in social reality and moral reality at the same time is reflected in calls for future research directions in artificial intelligence (e.g. Russell *et al.* [2015]) and visible in initial contributions towards that end (e.g. Conitzer *et al.* [2017]).

These general AI challenges provide a unique opportunity for the interdisciplinary field of normative multi-agent systems. This field studies the very dynamics that allow systems to address fuzzy and complex problems conventional rule-based systems are not prepared to deal with. It does so by exploiting two central features of norms, a) their adaptiveness towards changing social and technological environments, and b) their innate scalability based on their decentralized operation. Independent of the application domain, this leaves us researchers with the task to foster and establish an interdisciplinary operationalisation of norms as dynamic decentralized coordination mechanisms. This, in consequence, makes norm dynamics an integral component for the modelling of realistic social behaviour within and beyond normative multi-agent systems.

References

[Airiau *et al.*, 2014] S. Airiau, S. Sen, and D. Villatoro. Emergence of conventions through social learning. *Autonomous Agents and Multi-Agent Systems*, 28(5):779–804, 2014.

[Alchourrón and Makinson, 1981] C. E. Alchourrón and D. Makinson. Hierarchies of regulations and their logic. 1981. in [**?**] 125–148.

[Alchourrón and Makinson, 1982] C. E. Alchourrón and D. Makinson. On the logic of theory change: Contraction functions and their associated revision functions. *Theoria*, 48:14–37, 1982.

[Alchourrón *et al.*, 1985] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *J. Symb. Log.*, 50(2):510–530, 1985.

[Andreoni, 1989] J. Andreoni. Giving with impure altruism: Applications to charity and ricardian equivalence. *Journal of Political Economy*, 97(6):1447–1458, 1989.

[Andrighetto *et al.*, 2007] G. Andrighetto, M. Campenni, R. Conte, and M. Paolucci. On the immergence of norms: A normative agent architecture. In *Proceedings of the AAAI Symposium, Social and Organizational Aspects of Intelligence*, Washington DC, 2007.

[Andrighetto *et al.*, 2010] G. Andrighetto, M. Campenni, F. Cecconi, and R. Conte. The complex loop of norm emergence: A simulation model. In K. Takadama, C. Cioffi-Revilla, and G. Deffuant, editors, *Simulating Interacting Agents and Social Phenomena*, pages 19–35. Springer, Berlin, 2010.

[Andrighetto *et al.*, 2013] G. Andrighetto, C. Castelfranchi, E. Mayor, J. McBreen, M. Lopez-Sanchez, and S. Parsons. (Social) norm dynamics. In G. Andrighetto, G. Governatori, P. Noriega, and L. van der Torre, editors, *Normative Multi-Agent Systems. Vol. 4 of Dagstuhl Follow-Ups*, pages 135–170, 2013.

[Aucher *et al.*, 2009] G. Aucher, D. Grossi, A. Herzig, and E. Lorini. Dynamic context logic. In X. He, J. Horty, and E. Pacuit, editors, *Logic, Rationality, and Interaction: Second International Workshop, LORI 2009, Chongqing, China, October 8-11, 2009. Proceedings*, pages 15–26, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

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

[Baldwin, 1971] D. A. Baldwin. The power of positive sanctions. *World Politics*, 24(1):19Ű38, 1971.

[Bandura, 1977] A. Bandura. *Social Learning Theory*. General Learning Press, New York (NY), 1977.

[Barabási and Albert, 1999] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[Beheshti *et al.*, 2015] R. Beheshti, A. M. Ali, and G. Sukthankar. Cognitive social learners: An architecture for modeling normative behavior. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2017–2023. AAAI Press, 2015.

[Bianco and Bates, 1990] W. T. Bianco and R. Bates. Cooperation by design: Leadership, structure, and collective dilemmas. *American Political Science Review*, 84(1):133–147, 1990.

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

[Boella and van der Torre, 2004] G. Boella and L. van der Torre. Regulative and constitutive norms in normative multiagent systems. In *IN PROCS. OF KRÕ04*, pages 255–265. AAAI Press, 2004.

[Boella *et al.*, 2006] G. Boella, L. van der Torre, and H. Verhagen. Introduction to normative multiagent systems. *Computation and Mathematical Organizational Theory*, 12(2-3):71–79, 2006.

[Boella *et al.*, 2008] G. Boella, L. van der Torre, and H. Verhagen. Introduction to special issue on normative multiagent systems. *Autonomous Agents and Multi-Agent Systems*, 17(1):1–10, 2008.

[Boella *et al.*, 2009] G. Boella, G. Pigozzi, and L. van der Torre. Normative framework for normative system change. In *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, May 10-15, 2009, Volume 1*, pages 169–176, 2009.

[Boella *et al.*, 2016a] G. Boella, L. D. Caro, L. Humphreys, L. Robaldo, P. Rossi, and L. vanăder Torre. Eunomos, a legal document and knowledge management system for the web to provide relevant, reliable and up-to-date information on the law. *Artificial Intelligence and Law*, 24(3):245–283, 2016.

[Boella *et al.*, 2016b] G. Boella, G. Pigozzi, and L. van der Torre. AGM contraction and revision of rules. *Journal of Logic, Language and Information*, 25(3-4):273–297, 2016.

[Boman, 1999] M. Boman. Norms in artificial decision making. *Artificial Intelligence and Law*, 7(1):17–35, 1999.

[Bou *et al.*, 2007] E. Bou, M. López-Sánchez, and J. A. Rodríguez-Aguilar. Adaptation of autonomic electronic institutions through norms and institutional agents. In G. M. P. O'Hare, A. Ricci, M. J. O'Grady, and O. Dikenelli, editors, *Engineering Societies in the Agents World VII: 7th International Workshop, ESAW 2006 Dublin, Ireland, September 6-8, 2006 Revised Selected and Invited Papers*, pages 300–319, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[Bourdieu, 1977] P. Bourdieu. *An Outline of a Theory of Practice*. Cambridge University Press, London, 1977.

[Boyd and Richerson, 1985] R. Boyd and P. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, Chicago (IL), 1985.

[Boyd and Richerson, 2005] R. Boyd and P. Richerson. *The Origin and Evolution of Cultures*. Oxford University Press, New York (NY), 2005.

[Bratman, 1987] M. Bratman. *Intentions, Plans, and Practical Reason*. Harvard University Press, Cambridge (MA), 1987.

[Bravo *et al.*, 2012] G. Bravo, F. Squazzoni, and R. Boero. Trust and partner selection in social networks: An experimentally grounded model. *Social Networks*, 34(4):481–492, 2012.

[Broersen *et al.*, 2001] J. Broersen, M. Dastani, J. Hulstijn, Z. Huang, and L. van der Torre. The BOID architecture: Conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the Fifth International Conference on Autonomous Agents*, AGENTS '01, pages 9–16, New York, NY, USA, 2001. ACM.

[Broersen *et al.*, 2002] J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, 2(3-4):428–447, 2002.

[Caldas and Coelho, 1999] J. C. Caldas and H. Coelho. The origin of institutions: socio-economic processes, choice, norms and conventions. *Journal of Artificial Societies and Social Simulation*, 2(2):1, 1999.

[Campenni *et al.*, 2009] M. Campenni, G. Andrighetto, F. Cecconi, and R. Conte. Normal = Normative? The role of intelligent agents in norm innovation. *Mind & Society*, 8(2):153–172, 2009.

[Campos *et al.*, 2009] J. Campos, M. López-Sánchez, J. A. Rodríguez-Aguilar, and M. Esteva. Formalising situatedness and adaptation in electronic institutions. In J. F. Hübner, E. Matson, O. Boissier, and V. Dignum, editors, *Coordination, Organizations, Institutions and Norms in Agent Systems IV : COIN 2008 International Workshops, COIN@AAMAS 2008, Estoril, Portugal, May 12, 2008. COIN@AAAI 2008, Chicago, USA, July 14, 2008. Revised Selected Papers*, pages 126–139, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

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

[Chalub *et al.*, 2006] F. Chalub, F. Santos, and J. Pacheco. The evolution of norms. *Journal of Theoretical Biology*, 241(2):233–240, 2006.

[Checkel, 1998] J. Checkel. The constructivist turn in international relations theory. *World Politics*, 50(2):324–348, 1998.

[Christelis and Rovatsos, 2009] G. Christelis and M. Rovatsos. Automated norm synthesis in an agent-based planning environment. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '09, pages 161–168, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems.

[Conitzer *et al.*, 2017] V. Conitzer, W. Sinnott-Armstrong, J. S. Borg, Y. Deng, and M. Kramer. Moral decision making frameworks for artificial intelligence, 2017.

[Conte and Castelfranchi, 1995a] R. Conte and C. Castelfranchi. *Cognitive and Social Action*. UCL Press, London, 1995.

[Conte and Castelfranchi, 1995b] R. Conte and C. Castelfranchi. Understanding the effects of norms in social groups through simulation. In N. Gilbert and R. Conte, editors, *Artificial Societies: The Computer Simulation of Social Life*, pages 252–267. UCL Press, London, 1995.

[Crawford and Ostrom, 1995] S. E. Crawford and E. Ostrom. A Grammar of Institutions. *The American Political Science Review*, 89(3):582–600, September 1995.

[Delgado *et al.*, 2003] J. Delgado, J. M. Pujol, and R. Sangüesa. Emergence of coordination in scale-free networks. *Web Intelligence and Agent Systems*, 1(2):131–138, April 2003.

[Delgado, 2002] J. Delgado. Emergence of social conventions in complex networks. *Artificial Intelligence*, 141(1–2):171–185, 2002.

[Eguia, 2011] J. X. Eguia. *A Theory of Discrimination and Assimilation*. New York University Press, New York (NY), 2011.

[Ehrlich and Levin, 2005] P. R. Ehrlich and S. A. Levin. The evolution of norms. *PLoS Biology*, 3(6), 06 2005.

[Elster, 1989] J. Elster. Social norms and economic theory. *Journal of Economic Perspectives*, 3(4):99–117, 1989.

[Epstein, 2001] J. M. Epstein. Learning to be thoughtless: Social norms and individual computation. *Computational Economics*, 18(1):9–24, 2001.

[Erdős and Rényi, 1959] P. Erdős and A. Rényi. On random graphs i. *Publicationes Mathematicae*, 6:290–297, 1959.

[Fermé and Hansson, 1999] E. L. Fermé and S. O. Hansson. Selective revision. *Studia Logica: An International Journal for Symbolic Logic*, 63(3):331–342, 1999.

[Finnemore and Sikkink, 1998] M. Finnemore and K. Sikkink. International norm dynamics and political change. *International Organization*, 52(4):887–917, 1998.

[Fitoussi and Tennenholtz, 2000] D. Fitoussi and M. Tennenholtz. Choosing social laws for multi-agent systems: Minimality and simplicity. *Artificial Intelligence*, 119(1–2):61–101, 2000.

[Fix *et al.*, 2006] J. Fix, C. von Scheve, and D. Moldt. Emotion-based norm enforcement and maintenance in multi-agent systems: Foundations and Petri net modeling. In H. Nakashima, M. P. Wellman, G. Weiss, and P. Stone, editors, *Proceedings of the 5th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '06, pages 105–107, New York (NY), 2006. ACM Press.

[Flentge *et al.*, 2001] F. Flentge, D. Polani, and T. Uthmann. Modelling the emergence of possession norms using memes. *Journal of Artificial Societies and Social Simulation*, 4(4):3, 2001.

[Franks *et al.*, 2013] H. Franks, N. Griffiths, and S. Anand. Learning influence in complex social networks. In Ito, Jonker, Gini, and Shehory, editors, *Proceedings of the 12th International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '13, pages 447–454, 2013.

[Franks *et al.*, 2014] H. Franks, N. Griffiths, and S. S. Anand. Learning agent influence in MAS with complex social networks. *Autonomous Agents and Multi-Agent Systems*, 28:836–866, 2014.

[Frantz *et al.*, 2013] C. Frantz, M. K. Purvis, M. Nowostawski, and B. T. R. Savarimuthu. nADICO: A nested grammar of institutions. In G. Boella, E. Elkind, B. T. R. Savarimuthu, F. Dignum, and M. K. Purvis, editors, *PRIMA 2013: Principles and Practice of Multi-Agent Systems*, volume 8291 of *Lecture Notes in Artificial Intelligence*, pages 429–436, Berlin, 2013. Springer.

[Frantz *et al.*, 2014a] C. Frantz, M. K. Purvis, M. Nowostawski, and B. T. R. Savarimuthu. Modelling institutions using dynamic deontics. In T. Balke, F. Dignum, M. B. van Riemsdijk, and A. K. Chopra, editors, *Coordination, Organizations, Institutions and Norms in Agent Systems IX*, volume 8386 of *Lecture Notes in Artificial Intelligence*, pages 211–233, Berlin, 2014. Springer.

[Frantz *et al.*, 2014b] C. Frantz, M. K. Purvis, B. T. R. Savarimuthu, and M. Nowostawski. Analysing the dynamics of norm evolution using interval type-2 fuzzy sets. In *WI-IAT '14 Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 3, pages 230–237, 2014.

[Frantz *et al.*, 2014c] C. Frantz, M. K. Purvis, B. T. R. Savarimuthu, and M. Nowostawski. Modelling dynamic normative understanding in agent societies. In H. K. Dam, J. Pitt, Y. Xu, G. Governatori, and T. Ito, editors, *Principles and Practice of Multi-Agent Systems - 17th International Conference, PRIMA 2014*, volume 8861 of *Lecture Notes in Artificial Intelligence*, pages 294–310, Berlin, 2014. Springer.

[Frantz *et al.*, 2015] C. K. Frantz, M. K. Purvis, B. T. R. Savarimuthu, and M. Nowostawski. Modelling dynamic normative understanding in agent societies. *Scalable Computing: Practice and Experience*, 16(4):355–378, 2015.

[Frantz *et al.*, 2016] C. K. Frantz, B. T. R. Savarimuthu, M. K. Purvis, and M. Nowostawski. Generalising social structure using interval type-2 fuzzy sets. In M. Baldoni, A. K. Chopra, T. C. Son, K. Hirayama, and P. Torroni, editors, *PRIMA 2016: Principles and Practice of Multi-Agent Systems: 19th International Conference, Phuket, Thailand, August 22-26, 2016, Proceedings*, pages 344–354. Springer International Publishing, Cham, 2016.

[Frantz, 2015] C. K. Frantz. *Agent-Based Institutional Modelling: Novel Techniques for Deriving Structure from Behaviour.* PhD thesis, University of Otago, Dunedin, New Zealand, 2015. Available under: http://hdl.handle.net/10523/5906.

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

[Goette *et al.*, 2006] L. Goette, D. Huffman, and S. Meier. The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review*, 96(2):212–216, May 2006.

[Governatori and Rotolo, 2010] G. Governatori and A. Rotolo. Changing legal systems: legal abrogations and annulments in defeasible logic. *Logic Journal of IGPL*, 18(1):157–194, 2010.

[Governatori *et al.*, 2013] G. Governatori, A. Rotolo, F. Olivieri, and S. Scannapieco. Legal contractions: A logical analysis. In E. Francesconi and B. Verheij, editors, *ICAIL*, pages 63–72. ACM, 2013.

[Greenwald *et al.*, 2002] A. G. Greenwald, M. R. Banaji, L. A. Rudman, S. D. Farnham, B. A. Nosek, and D. S. Mellott. A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109(1):3–25, 2002.

[Griffiths and Luck, 2010] N. Griffiths and M. Luck. Norm diversity and emergence in tag-based cooperation. In M. D. Vos, N. Fornara, J. V. Pitt, and G. A. Vouros, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems VI - COIN 2010 International Workshops, COIN@AAMAS 2010, Toronto, Canada, May 2010, COIN@MALLOW 2010, Lyon, France, Au-*

gust 2010, Revised Selected Papers, volume 6541 of *Lecture Notes in Computer Science*, pages 230–249. Springer, 2010.

[Grossi *et al.*, 2008] D. Grossi, J.-J. Meyer, and F. Dignum. The many faces of counts-as: A formal analysis of constitutive-rules. *J. of Applied Logic*, 6(2):192–217, 2008.

[Hales, 2002] D. Hales. Group reputation supports beneficent norms. *Journal of Artificial Societies and Social Simulation*, 5(4):4, 2002.

[Hansen *et al.*, 2007] J. Hansen, G. Pigozzi, and L. van der Torre. Ten philosophical problems in deontic logic. In G. Boella, L. van der Torre, and H. Verhagen, editors, *Normative Multi-Agent Systems. Dagstuhl Seminar Proc. 07122*, 2007.

[Hansson, 1993] S. Hansson. Reversing the Levi identity. *Journal of Philosophical Logic*, 22:637–669, 1993.

[Henderson, 2005] D. Henderson. Norms, invariance, and explanatory relevance. *Philosophy of the Social Sciences*, 35(3):324–338, 2005.

[Hoffmann, 2003] M. Hoffmann. Entrepreneurs and norm dynamics: An agent-based model of the norm life cycle. Technical report, Department of Political Science and International Relations, University of Delaware, Newark (DE), 2003.

[Hoffmann, 2005] M. Hoffmann. Self-organized criticality and norm avalanches. In *Proceedings of the Symposium on Normative Multi-Agent Systems*, Hatfield (UK), 2005. AISB.

[Hogg, 2001] M. A. Hogg. A social identity theory of leadership. *Personality and Social Psychology Review*, 5(3):184–200, 2001.

[Hollander and Wu, 2011a] C. Hollander and A. Wu. Using the process of norm emergence to model consensus formation. In *Fifth IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO)*, pages 148–157, Oct 2011.

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

[Horne, 2001] C. Horne. Sociological perspectives on the emergence of norms. In M. Hechter and K. Opp, editors, *Social Norms*, pages 3–34. Russell Sage Foundation, New York (NY), 2001.

[Horne, 2007] C. Horne. Explaining norm enforcement. *Rationality and Society*, 19(2):139–170, 2007.

[Huang *et al.*, 2016] X. Huang, J. Ruan, Q. Chen, and K. Su. Normative multiagent systems: A dynamic generalization. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1123–1129. AAAI Press, 2016.

[Huber, 1999] M. J. Huber. JAM: A BDI-theoretic mobile agent architecture. In O. Etzioni, J. P. Müller, and J. M. Bradshaw, editors, *Proceedings of the Third International Conference on Autonomous Agents (Agents '99)*, pages 236–243, Seattle (WA), 1999.

[Kahneman and Tversky, 1972] D. Kahneman and A. Tversky. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3):430 – 454, 1972.

[Kittock, 1995] J. Kittock. 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*, pages 507–521. Addison-Wesley, 1995.

[Lewis, 1969] D. K. Lewis. *Convention: A Philosophical Study*. Harvard University Press, Cambridge (MA), 1969.

[López y López and Luck, 2004] F. López y López and M. Luck. Towards a model of the dynamics of normative multi-agent systems. In G. Lindemann, D. Moldt, and M. Paolucci, editors, *RASTA 2002*, volume 2934 of *Lecture Notes in Artificial Intelligence*, pages 175–194, Heidelberg, 2004. Springer.

[López y López and Márquez, 2004] F. López y López and A. A. Márquez. An architecture for autonomous normative agents. In *Proceedings of the Fifth Mexican International Conference in Computer Science - ENC*, pages 96–103, Los Alamitos, CA, USA, 2004. IEEE Computer Society.

[López y López *et al.*, 2002] F. López y López, M. Luck, and M. d'Inverno. Constraining autonomy through norms. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems AAMAS*, pages 674–681, New York, NY, USA, 2002. ACM.

[López y López *et al.*, 2006] F. López y López, M. Luck, and M. d'Inverno. A normative framework for agent-based systems. *Computational & Mathematical Organization Theory*, 12(2):227–250, 2006.

[López y López *et al.*, 2007] F. López y López, M. Luck, and M. d'Inverno. A normative framework for agent-based systems. In G. Boella, L. van der Torre, and H. Verhagen, editors, *Normative Multi-agent Systems, Dagstuhl Seminar Proceedings 07122*, Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2007.

[López y López, 2003] F. López y López. *Social Powers and Norms: Impact on Agent Behaviour*. PhD thesis, Department of Electronics and Computer Science, University of Southampton, United Kingdom, 2003.

[Mahmoud *et al.*, 2012] S. Mahmoud, N. Griffiths, J. Keppens, and M. Luck. Efficient norm emergence through experiential dynamic punishment. In *ECAI'12*, pages 576–581, 2012.

[Mahmoud *et al.*, 2014a] M. A. Mahmoud, M. S. Ahmad, M. Z. M. Yusoff, and A. Mustapha. Norms assimilation in heterogeneous agent community. In H. K. Dam, J. Pitt, Y. Xu, G. Governatori, and T. Ito, editors, *Principles and Practice of Multi-Agent Systems - 17th International Conference, PRIMA 2014*, volume 8861 of *Lecture Notes in Artificial Intelligence*, pages 311–318, Berlin, 2014. Springer.

[Mahmoud *et al.*, 2014b] M. A. Mahmoud, M. S. Ahmad, M. Z. M. Yusoff, and A. Mustapha. A review of norms and normative multiagent systems. *The Scientific World Journal*, 2014:23 pages, 2014. Article ID 684587.

[Mahmoud *et al.*, 2015] S. Mahmoud, N. Griffiths, J. Keppens, A. Taweel, T. J. Bench-Capon, and M. Luck. Establishing norms with metanorms in distributed computational systems. *Artif. Intell. Law*, 23(4):367–407, December 2015.

[Makinson and van der Torre, 2000] D. Makinson and L. van der Torre. Input/output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.

[Makinson and van der Torre, 2003] D. Makinson and L. van der Torre. What is input/output logic. In B. Löwe, W. Malzkom, and T. Räsch, editors, *Foundations of the Formal Sciences II : Applications of Mathematical Logic in Philosophy and Linguistics (Papers of a conference held in Bonn, November 10-13, 2000)*, Trends in Logic, vol. 17, pages 163–174, Dordrecht, 2003. Kluwer.

Reprinted in this volume.

[Manna and Waldinger, 1980] Z. Manna and R. Waldinger. A deductive approach to program synthesis. *ACM Transactions on Programming Languages and Systems*, 2(1):90–121, 1980.

[Maranhão, 2001] J. Maranhão. Refinement. A tool to deal with inconsistencies. In *Proceedings of the 8th ICAIL*, pages 52–59, 2001.

[Maranhão, 2017] J. Maranhão. A logical architecture for dynamic legal interpretation. In *Proceedings of 16th ICAIL*, pages 129–139, 2017.

[Meneguzzi and Luck, 2009] F. Meneguzzi and M. Luck. Norm-based behaviour modification in bdi agents. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '09, pages 177–184, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems.

[Mihaylov *et al.*, 2014] M. Mihaylov, K. Tuyls, and A. Nowé. A decentralized approach for convention emergence in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 28(5):749–778, 2014.

[Morales *et al.*, 2013] J. Morales, M. López-Sánchez, J. A. Rodríguez-Aguilar, M. Wooldridge, and W. Vasconcelos. Automated synthesis of normative systems. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '13, pages 483–490, 2013.

[Morales *et al.*, 2014] J. Morales, M. López-Sánchez, J. A. Rodríguez-Aguilar, M. Wooldridge, and W. Vasconcelos. Minimality and simplicity in the on-line automated synthesis of normative systems. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '14, pages 109–116, 2014.

[Morales *et al.*, 2015a] J. Morales, M. López-Sánchez, J. A. Rodriguez-Aguilar, W. Vasconcelos, and M. Wooldridge. Online automated synthesis of compact normative systems. *ACM Trans. Auton. Adapt. Syst.*, 10(1):2:1–2:33, March 2015.

[Morales *et al.*, 2015b] J. Morales, M. López-Sánchez, J. A. Rodríguez-Aguilar, M. Wooldridge, and W. Vasconcelos. Synthesising liberal normative systems. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pages 433–441, Richland, SC, 2015. International Foundation for Autonomous Agents and Multiagent Systems.

[Morales *et al.*, 2015c] J. Morales, I. Mendizabal, D. Sanchez-Pinsach, M. López-Sánchez, and J. A. Rodríguez-Aguilar. Using IRON to build frictionless on-line communities. *AI Commun.*, 28(1):55–71, 2015.

[Mukherjee *et al.*, 2007] P. Mukherjee, S. Sen, and S. Airiau. Emergence of norms with biased interaction in heterogeneous agent societies. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 512–515, 2007.

[Mukherjee *et al.*, 2008] P. Mukherjee, S. Sen, and S. Airiau. Norm emergence under constrained interactions in diverse societies. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '08, pages 779–786, Richland, SC, 2008. International Foundation for Autonomous Agents and Multiagent Systems.

[Nadelman, 1990] E. Nadelman. Global prohibition regimes: The evolution of norms in international society. *International Organization*, 44(4):479–526, 1990.

[Nakamaru and Levin, 2004] M. Nakamaru and S. A. Levin. Spread of two linked social norms on complex interaction networks. *Journal of Theoretical Biology*, 230(1):57–64, 2004.

[Onn and Tennenholtz, 1997] S. Onn and M. Tennenholtz. Determination of social laws for multi-agent mobilization. *Artificial Intelligence*, 95(1):155–167, 1997.

[Ossowski, 2013] S. Ossowski. *Agreement Technologies*. Springer, Dordrecht (NL), 2013.

[Perreau de Pinninck *et al.*, 2008] A. Perreau de Pinninck, C. Sierra, and M. Schorlemmer. Distributed norm enforcement via ostracism. In J. Sichman, J. Padget, S. Ossowski, and P. Noriega, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems III*, volume 4870 of *Lecture Notes in Computer Science*, pages 301–315. Springer, Berlin, 2008.

[Perreau de Pinninck *et al.*, 2010] A. Perreau de Pinninck, C. Sierra, and M. Schorlemmer. A multiagent network for peer norm enforcement. *Autonomous Agents and Multi-Agent Systems*, 21(3):397–424, 2010.

[Pigozzi and van der Torre, 2017] G. Pigozzi and L. van der Torre. Multiagent deontic logic and its challenges from a normative systems perspective. *The IfCoLog Journal of Logics and their Applications*, 4(9), 2017.

[Pucella and Weissman, 2004] R. Pucella and V. Weissman. Reasoning about dynamic policies. In I. Walukiewicz, editor, *FOSSACS 2004. LNCS, vol. 2987*, pages 453–467. Springer, Heidelberg, 2004.

[Pujol *et al.*, 2005] J. M. Pujol, J. Delgado, R. Sangüesa, and A. Flache. The role of clustering on the emergence of efficient social conventions. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 965–970, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[Rao and Georgeff, 1995] A. S. Rao and M. P. Georgeff. BDI agents: From theory to practice. *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*, pages 312–319, 1995.

[Risse and Sikkink, 1999] S. R. Risse, Thomas and K. Sikkink. *The Power of Human Rights: International Norms and Domestic Change*. Cambridge University Press, Cambridge, 1999.

[Riveret *et al.*, 2012] R. Riveret, A. Rotolo, and G. Sartor. Probabilistic rule-based argumentation for norm-governed behaviour. *Artificial Intelligence*, 20(4):383–420, 2012.

[Riveret *et al.*, 2013] R. Riveret, G. Contissa, D. Busquets, A. Rotolo, J. Pitt, and G. Sartor. Vicarious reinforcement and ex ante law enforcement: A study in norm-governed learning agents. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ICAIL '13, pages 222–226, New York, NY, USA, 2013. ACM.

[Riveret *et al.*, 2014] R. Riveret, A. Artikis, D. Busquets, and J. Pitt. Self-governance by transfiguration: From learning to prescriptions. In F. Cariani, D. Grossi, J. Meheus, and X. Parent, editors, *Deontic Logic and Normative Systems*, volume 8554 of *Lecture Notes in Computer Science*, pages 177–191. Springer, Springer, 2014.

[Russell *et al.*, 2015] S. Russell, D. Dewey, and M. Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114, 2015.

[Saam and Harrer, 1999] N. J. Saam and A. Harrer. Simulating norms, social inequality, and functional change in artificial societies. *Journal of Artificial Societies and Social Simulation*, 2(1):2, 1999.

[Salazar *et al.*, 2010] N. Salazar, J. A. Rodriguez-Aguilar, and J. L. Arcos. Robust coordination in large convention spaces. *AI Communications*, 23(4):357–372, 2010.

[Savarimuthu *et al.*, 2011] B. Savarimuthu, R. Arulanandam, and M. Purvis. Aspects of active norm learning and the effect of lying on norm emergence in agent societies. In D. Kinny, J.-j. Hsu, G. Governatori, and A. Ghose, editors, *Agents in Principle, Agents in Practice*, volume 7047 of *Lecture Notes in Computer Science*, pages 36–50. Springer, Berlin, 2011.

[Savarimuthu and Cranefield, 2009] B. T. R. Savarimuthu and S. Cranefield. A categorization of simulation works on norms. In G. Boella, G. Pigozzi, and L. van der Torre, editors, *Normative Multi-agent Systems, Dagstuhl Seminar Proceedings 09121*, pages 39–58, Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2009.

[Savarimuthu and Cranefield, 2011] B. T. R. Savarimuthu and S. Cranefield. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems*, 7(1):21–54, January 2011.

[Savarimuthu *et al.*, 2007] B. T. R. Savarimuthu, S. Cranefield, M. Purvis, and M. Purvis. Norm emergence in agent societies formed by dynamically changing networks. In *2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 464–470, 2007.

[Savarimuthu *et al.*, 2008a] B. T. R. Savarimuthu, S. Cranefield, M. Purvis, and M. Purvis. Role model based mechanism for norm emergence in artificial agent societies. In J. Sichman, J. Padget, S. Ossowski, and P. Noriega, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems III*, volume 4870 of *Lecture Notes in Computer Science*, pages 203–217. Springer, Berlin, 2008.

[Savarimuthu *et al.*, 2008b] B. T. R. Savarimuthu, M. A. Purvis, and M. K. Purvis. Social norm emergence in virtual agent societies. In *Proceedings of the 7th International Conference on Autonomous Agents and Multi-Agent Systems*, 2008.

[Savarimuthu *et al.*, 2009] B. T. R. Savarimuthu, S. Cranefield, M. A. Purvis, and M. K. Purvis. Norm emergence in agent societies formed by dynamically changing networks. *Web Intelligence and Agent Systems*, 7(3):223–232, 2009.

[Savarimuthu *et al.*, 2010a] B. T. R. Savarimuthu, S. Cranefield, M. A. Purvis, and M. K. Purvis. A data mining approach to identify obligation norms in agent societies. In *Proceedings of the International Workshop on Agents and Data Mining Interaction (ADMI@AAMAS 2010), Toronto, Canada*, pages 54–69, May 2010.

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

[Savarimuthu *et al.*, 2013a] B. T. R. Savarimuthu, S. Cranefield, M. A. Purvis, and M. K. Purvis. Identifying prohibition norms in agent societies. *Artificial Intelligence and Law*, 21:1–46, 2013.

[Savarimuthu *et al.*, 2013b] B. T. R. Savarimuthu, J. Padget, and M. A. Purvis. Social norm recommendation for virtual agent societies. In G. Boella, E. Elkind, B. T. R. Savarimuthu, F. Dignum, and M. K. Purvis, editors, *PRIMA 2013: Principles and Practice of Multi-Agent Systems*, volume

8291 of *Lecture Notes in Computer Science*, pages 308–323. Springer, Berlin, 2013.

[Savarimuthu, 2011] B. T. R. Savarimuthu. *Mechanisms for Norm Emergence and Norm Identification in Multi-Agent Societies*. PhD thesis, University of Otago, Dunedin, New Zealand, 2011.

[Scheve *et al.*, 2006] C. v. Scheve, D. Moldt, J. Fix, and R. v. Luede. My agents love to conform: Norms and emotion in the micro-macro link. *Computational & Mathematical Organization Theory*, 12(2):81–100, 2006.

[Schneider and Teske, 1992] M. Schneider and P. Teske. Toward a theory of the political entrepreneur: Evidence from local government. *American Political Science Review*, 86(3):737–747, 1992.

[Sen and Airiau, 2007] S. Sen and S. Airiau. Emergence of norms through social learning. In M. Veloso, editor, *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1507–1512, San Francisco (CA), 2007. Morgan Kaufmann Publishers Inc.

[Sen and Sen, 2010] O. Sen and S. Sen. Effects of social network topology and options on norm emergence. In J. Padget, A. Artikis, W. Vasconcelos, K. Stathis, V. da Silva, E. Matson, and A. Polleres, editors, *Coordination, Organizations, Institutions and Norms in Agent Systems V*, volume 6069 of *Lecture Notes in Computer Science*, pages 211–222. Springer, Berlin, 2010.

[Shoham and Tennenholtz, 1992a] J. Shoham and M. Tennenholtz. Emergent conventions in multi-agent systems: Initial experimental results and observations. In *Proceedings of the Third International Conference on the Principles of Knowledge Representation and Reasoning KR*, pages 225–231, San Mateo, CA, USA, 1992. Morgan Kaufmann.

[Shoham and Tennenholtz, 1992b] Y. Shoham and M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies. In *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI '92)*, pages 276–281, San Jose (CA), July 1992.

[Shoham and Tennenholtz, 1995] Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73(1û2):231 – 252, 1995. Computational Research on Interaction and Agency, Part 2.

[Shoham and Tennenholtz, 1997] Y. Shoham and M. Tennenholtz. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94(1–2):139–166, July 1997.

[Simon, 1955] H. A. Simon. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955.

[Sims and Brinkmann, 2003] R. R. Sims and J. Brinkmann. Enron ethics (or: Culture matters more than codes). *Journal of Business Ethics*, 45(3):243–256, Jul 2003.

[Singh, 2014] M. P. Singh. Norms as a basis for governing sociotechnical systems. *ACM Trans. Intell. Syst. Technol.*, 5(1):21:1–21:23, January 2014.

[Staller and Petta, 2001] A. Staller and P. Petta. Introducing emotions into the computational study of social norms: A first evaluation. *Journal of Artificial Societies and Social Simulation*, 4(1):2, 2001.

[Stolpe, 2010] A. Stolpe. Norm-system revision: Theory and application. *Artificial Intelligence and Law*, 18:247–283, 2010.

[Sugawara, 2011] T. Sugawara. Emergence and stability of social conventions in conflict situations. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume One*, IJCAI'11, pages 371–378. AAAI Press, 2011.

[Sunstein, 1996] C. R. Sunstein. Social norms and social roles. *Columbia Law Review*, 96(4):903–968, 1996.

[Tinnemeier *et al.*, 2009] N. Tinnemeier, M. Dastani, J.-J. C. Meyer, and L. van der Torre. Programming normative artifacts with declarative obligations and prohibitions. In *Proc. of WI/IATÕ09*. IEEE Computer Society, 2009.

[Tinnemeier *et al.*, 2010] N. Tinnemeier, M. Dastani, and J.-J. Meyer. Programming norm change. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10, pages 957–964, Richland, SC, 2010. International Foundation for Autonomous Agents and Multiagent Systems.

[Urbano *et al.*, 2009] P. Urbano, J. Balsa, L. Antunes, and L. Moniz. Force versus majority: A comparison in convention emergence efficiency. In J. Hübner, E. Matson, O. Boissier, and V. Dignum, editors, *Coordination, Organizations, Institutions and Norms in Agent Systems IV*, volume 5428 of *Lecture Notes in Computer Science*, pages 48–63. Springer, Berlin, 2009.

[Valk, 1998] R. Valk. Petri nets as token objects. an introduction to elementary. In *Proceedings of Application and Theory of Petri Nets*, pages 1–25, Berlin, 1998. Springer.

[van Ditmarsch and van der Hoek, 2007] H. van Ditmarsch and K. B. van der Hoek, W. *Dynamic Epistemic Logic*. Synthese Library Series, vol. 337, Springer, Heidelberg, 2007.

[Verhagen, 2000] H. J. Verhagen. *Norm Autonomous Agents*. PhD thesis, The Royal Institute of Technology and Stockholm University, Stockholm, Sweden, 2000.

[Verhagen, 2001] H. Verhagen. Simulation of the learning of norms. *Social Science Computer Review*, 19(3):296–306, 2001.

[Villatoro *et al.*, 2009] D. Villatoro, S. Sen, and J. Sabater-Mir. 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 - Volume 02*, WI-IAT '09, pages 233–240, 2009.

[Villatoro *et al.*, 2011a] D. Villatoro, G. Andrighetto, J. Sabater-Mir, and R. Conte. Dynamic sanctioning for robust and cost-efficient norm compliance. In T. Walsh, editor, *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, volume 1 of *IJCAI'11*, pages 414–419. AAAI Press, 2011.

[Villatoro *et al.*, 2011b] D. Villatoro, J. Sabater-Mir, and S. Sen. Social instruments for robust convention emergence. In T. Walsh, editor, *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, volume 1 of *IJCAI'11*, pages 420–425. AAAI Press, 2011.

[Villatoro *et al.*, 2013] D. Villatoro, J. Sabater-Mir, and S. Sen. Robust convention emergence in social networks through self-reinforcing structures dissolution. *ACM Trans. Auton. Adapt. Syst.*, 8(1):2:1–2:21, April 2013.

[Walker and Wooldridge, 1995] A. Walker and M. Wooldridge. Understanding the emergence of conventions in multi-agent systems. In *Proceedings of the first international conference on multiagent systems (ICMAS)*, pages 384–389, Menlo Park (CA), 1995. AAAI Press.

[Watkins and Dayan, 1992] C. Watkins and P. Dayan. Technical note: Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

[Watts and Strogatz, 1998] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[Young, 1990] O. Young. Political leadership and regime formation: On the development of institutions in international society. *International Organization*, 45(3):281–308, 1990.

[Younger, 2004] S. Younger. Reciprocity, normative reputation, and the development of mutual obligation in gift-giving societies. *Journal of Artificial Societies and Social Simulation*, 7(1):5, 2004.

[Yu *et al.*, 2010] C.-H. Yu, J. Werfel, and R. Nagpal. Collective decision-making in multi-agent systems by implicit leadership. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 3 - Volume 3*, AAMAS '10, pages 1189–1196, Richland, SC, 2010. International Foundation for Autonomous Agents and Multiagent Systems.

[Yu *et al.*, 2013] C. Yu, M. Zhang, F. Ren, and X. Luo. Emergence of social norms through collective learning in networked agent societies. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '13, pages 475–482, 2013.

[Yu *et al.*, 2015] C. Yu, H. Lv, F. Ren, H. Bao, and J. Hao. Hierarchical learning for emergence of social norms in networked multiagent systems. In *AI 2015: Advances in Artificial Intelligence: 28th Australasian Joint Conference, Canberra, ACT, Australia, November 30 – December 4, 2015, Proceedings*, pages 630–643, Cham, 2015. Springer International Publishing.

[Zhang and Leezer, 2009] Y. Zhang and J. Leezer. Emergence of social norms in complex networks. In *International Conference on Computational Science and Engineering*, pages 549–555, Vancouver, 2009.

Received 5 September 2016