Social Coordination: Principles, Artefacts and Theories (SOCIAL.PATH)

Harko Verhagen, Pablo Noriega, Tina Balke and Marina de Vos (editors)

AISB Convention 2013 • University of Exeter • 3rd–5th April, 2013

Foreword from the Convention Chairs

This volume forms the proceedings of one of eight co-located symposia held at the AISB Convention 2013 that took place 3rd-5th April 2013 at the University of Exeter, UK. The convention consisted of these symposia together in four parallel tracks with five plenary talks; all papers other than the plenaries were given as talks within the symposia. This symposium-based format, which has been the standard for AISB conventions for many years, encourages collaboration and discussion among a wide variety of disciplines. Although each symposium is self contained, the convention as a whole represents a diverse array of topics from philosophy, psychology, computer science and cognitive science under the common umbrella of artificial intelligence and the simulation of behaviour.

We would like to thank the symposium organisers and their programme committees for their hard work in publicising their symposium, attracting and reviewing submissions and compiling this volume. Without these interesting, high quality symposia the convention would not be possible.

Dr Ed Keedwell & Prof. Richard Everson
AISB 2013 Convention Chairs

Published by
The Society for the Study of Artificial Intelligence and the Simulation of Behaviour
http://www.aisb.org.uk

ISBN: 978-1-908187-36-9

Table of Contents

Social Coordination: Principles, Artifacts and Theories

Tina Balke[1] and **Pablo Noriega**[2] and **Harko Verhagen**[3] and **Marina De Vos**[4]

Preface

Social science concepts such as norms, markets and rationality have found their way into computer science in general and agent-based research in particular where they model coordination and cooperation between largely independent autonomous computational entities. Vice versa, in the social sciences - sociology, philosophy, economics, legal science, etc. - computational models and their implementations have been used to investigate the rigour of theories and hypothesis. The use of these social science concepts in computer science is sometimes on a more metaphorical level than a detailed implementation of the ”real” concept and the theories surrounding it. Equally, the computer models used in the social sciences are not always convincing.

After a history of around 30 years of agents in computer science, the meeting of these two worlds is long overdue. This symposium aims to provide a meeting point for members of these communities to converse on principles, theories and artefacts for social coordination with the aim of facilitating future interactions and research.

The theme of social coordination was chosen because convergence is evident in this area.

This three-day workshop aims to (i) outline a shared perspective on the field (ii) identify challenges and opportunities for cross-disciplinary collaboration (iii) provide guidelines for research and policy-making and (iv) kindle partnerships among participants.

To achieve these objectives, we invited submisison of position papers, provocative papers, early, on-going and mature research papers, works discussing opportunities for interdisciplinary collaboration or policy guidelines and recommendations. We received 19 papers of which we selected 17.

These papers will give a short presentation slot, leaving ample time for discussion. Supplementing these presentations, we invited three invited speakers: Prof. Nigel Gilbert, from the University of Surrey providing his view from the social sciences, Prof. Michael Luck, King’s College London, offering a computer science perspective and Prof. Cristiano Castelfranchi, Institute of Reseach Psychology of Italian National Research council, will give his view on combining both disciplines.

We hope you enjoy the workshop and are looking forward to your views,

Tina, Pablo, Harko, Marina

[1] CRESS, University of Surrey, UK, email: t.balke@surrey.ac.uk
[2] IIIA-CSIC, Spain, email: pablo@iiia.csic.es
[3] Stockholm University, Sweden, email: verhagen@dsv.su.se
[4] University of Bath, UK, email: mdv@cs.bath.ac.uk

Social Mind As Coordination Artifact

Cristiano Castelfranchi [1]

Why human intelligence is social in its origin and function? We are (quite) intelligent, autonomous, self-motivated agents but in a "common world": dependence, strategic interaction, power.

But: What does "intelligence" mean? What does "social" mean?

"Social" doesn't mean "collective": social actions, social powers, social beliefs and emotions, ... are at the individual action and mind level; for (positive or negative) interactions. This social layer of mind and action is the presupposition and ground of "collective" behaviors, intelligence, institutions, organizations. No collective tricks without social minds.

"Social" doesn't mean pro-social, cooperative. Also exploitation and domination and subjection are social; conflicts and struggles are social interactions; and also conflicts requires "coordination". [And"cooperation" is not equal to "exchange"].

"*Coordination*" doesn't mean that X predicts the behavior of Y and the possible "interferences" with his own behaviors/goals to be avoided or exploited, and thus adjusts his own behavior to Y's expected behavior; and Y does the same. "Coordination" means also to *intentionally adjust (modify) the behavior of the other for our convenience*, by modifying his beliefs, emotions, and goals. We coordinate with the others by "influencing" them. Social power over the others is also for that. "Theory of Mind" and "mind reading" ability are for that; not just for predicting and explaining but for changing those behaviors/minds. "Coordination" and mind reading and mutual understanding, do not require language; they can be based on other forms of communication (behavioral and stigmergic: coordination artifacts, like language); or on no communication at all, just signs/cues reading.

Collective coordinated actions/behaviors do not necessarily require shared beliefs or goals, collective mind, joint action, acceptance, ... *There is efficient cooperation without joint mental plans, and even without plans at all.* Coordination and cooperation can just emerge by self-organization. This is the problem: *What is the relation between Emergence and Cognition? Between what we understand and intend and what we (collectively) do?*

A) The teleology of "functions" is not the teleology of intentions. *Functional*, adaptive social behaviors versus *mind-based*, *goal-directed*, *intelligent*, social actions. In humans and Artificial systems we have both:

   a) S-R reactive behaviors, just selected or designed or conditioned;

   b) Goal-directed (and intentional) social actions and social structures;

   c) self-organizing, emerging social orders, and (unintended) "social functions" over our habitual but also intentional behaviors.

   Social order as unintended, and social power as "alienation". So, the outcome of the "spontaneous" (Hayek) organization of society and of the "invisible hand" (A. Smith) is not the best possible or a good result. We let emerge and we (unintentionally) reproduce some equilibrium that can be very bad for the large majority of people, reflecting the inequality of powers and/or our ignorance and impotence.

B) We have to make explicit the "mental mediators" of social phenomena / objects (like norms, organization, power, institutions, money, ...). What the agents have to have in mind for producing that collective/sociological phenomenon? Coordination is through mind; not necessarily through "sharing".

In this perspective "social" (in the sense of "socially shaped") minds *are* "coordination artifacts". Cognitive and externalized Coordination Artifacts:

- cultural technical artifacts embodied in our mind and mental processing: language, time, computing, .. ;
- of special interest the "deontic" tricks: conventions, social norms, permissions, agreements (commitment and trust), legal norms and authorities, ... ; common values,..;
- "institutional" effects, actions, and roles;
- knowledge as collective institution; script and frames; education and instruction;...

[1] GOAL lab ISTC-CNR, Rome

Coordination: why is it so difficult to model?

Nigel Gilbert [1]

Abstract. Coordination is a basic feature of human interaction. Indeed, failure to achieve coordination might expose you ridicule, censure, or concerns about your mental capacity. On the other hand, human coordination is fragile, difficult to achieve, and often short-lived. At first sight, it may seem an easy task to engineer coordination in artificial societies to match that found in human ones. But, as I hope to show in this talk, understanding and modelling coordination is much more difficult than it seems. There are some fundamental problems in identifying coordination that computational systems tend at the moment to sidestep.

I start from a definition of coordination that it is 'the state of working together' (definition derived from http://en.wiktionary.org/wiki/coordination)

The first problem is to identify instances of coordination. Simply observing the behaviours of two actors that take place simultaneously is not sufficient: we don't know from observation of physical events that there is coordination involved. The behaviours may be the same, but performed independently and even without knowledge of the other. Or the behaviours may be different, yet coordinated.

The second problem is that even when we are told or come to understand what is going on, we still don't know whether the actions are really coordinated, especially if the intended coordination fails. This is because humans, unlike some computational agents, are not able to look inside the cognitions of their colleagues.

The third problem is that coordination may be on the basis of cooperation (the case we usually think of), or competition. The Prisoner's Dilemma is a good case of where it is not so clear whether coordination exists, in a context of cooperation and competition.

The fourth problem is that it may seem that coordination is best identified by asking the actors involved, but they face much the same kinds of difficulty as observers do in attributing intentionality to their allegedly coordinating partners or even to their own behaviours.

The fifth problem is that 'working together' implies some rules (or norms) about what to do, yet there are difficulties in applying rules to action that make the idea of 'rule-governed action' deeply problematic.

In this talk, I shall illustrate these problems with a real example, and begin to suggest what might be involved in designing a computational model that respects the limitations faced by human actors in attempting to achieve coordination.

ACKNOWLEDGEMENTS

This work was partially supported by the FP7-ICT Science of Global Systems programme of the European Commission through projects GLODERS (grant 315874) and ePolicy (grant 288147). It reects solely the views of its authors. The European Commission is not liable for any use that may be made of the information in the work.

[1] University of Surrey, England, email: n.gilbert@surrey.ac.uk

Towards Electronic Order

Michael Luck [1]

1 INTRODUCTION

A dictionary definition suggests that to *coordinate* is to “bring (parts, movements, etc.) into proper relation, cause to function together or in proper order.” This applies without consideration of whether the parts or movements are of mind, of human, of machine or of software. Perhaps this is so as a result of the definition preceding what is now common currency (certainly in our research communities, but increasingly so more generally), the idea of coordination of computational systems, with a firm footing in multi-agent systems and related areas.

The same dictionary defines *social* as “living in companies or organised communities, gregarious; . . . interdependent, co-operative, practising division of labour; existing only as member of compound organism, . . . ,” but also as “ concerned with the mutual relations of (classes of) human beings . . . ” The definition also includes reference to social animals but focusses on such natural systems and omits any consideration of artificial constructions. Again, this may be due to matters of time and technological developments.

While the concept of coordination might have emerged from human interactions, and from consideration in social sciences, to suggest that “use of these social science concepts in computer science is sometimes merely at a more metaphorical level than a detailed implementation of the *real* concept and the theories surrounding it,” offers only a narrow and traditional perspective. Sociality, coordination, and their combination as social coordination, are concepts that not only have value in application to artificial systems as well as natural systems, but also are manifest in artificial systems in ways that are as real and genuine as any other.

This paper argues against the notion of social coordination as metaphor in computational systems, and suggests instead that computational systems offer new insights into, techniques for, and realisations of both existing instantiations of social coordination and complex novel approaches to social coordination that extend the reach of the concept and our understanding of it. While some such computational techniques may be inspired and informed by social coordination in natural systems, in many cases the techniques are more sophisticated and complex than can easily be accommodated in natural systems. Yet this should not exclude them from our understanding of the concept; it might instead be considered as a paradigm shift in understanding a phenomenon similar to those encountered in scientific revolutions as suggested by Kuhn [3], for example.

2 SOCIAL COORDINATION

Taking social coordination as an umbrella concept that, in addition to traditional techniques (such as the contract net protocol) for providing order to enable effective distributed computing systems, also covers such aspects as norms, trust and reputation, we should be able to see that this is fundamental not just to future visions of computing, but to computing as it is today. Certainly, it is crucial if our distributed computing infrastructure is to be effective. However, the starting point for considering this is not in the social, but in the individual, where we must be concerned with not just what to do, but also how to do it. *Motivations*, which provide the *reasons* for doing something, also offer *constraints* on how a goal might best be achieved when faced with alternative courses of action [6].

In this context, and as discussed elsewhere [7, 5], motivations characterise the nature of agents: at extreme points, whether they are malevolent or benign. Thus agents may be well integrated members of a system or society, cooperating with others when requested to do so, participating effectively in joint ventures, and contributing to the good of the whole. Alternatively, they may be malicious, seeking to take advantage of others, by requesting cooperation but not providing it, by taking the benefits provided by a society without contributing to its success, or they may simply be incompetent or unable to deliver.

Various mechanisms may be required to ensure effective system operation. For example, in cases where there is a prevalence of benign behaviour from individual agents, there is less risk in interacting with agents because they will generally seek to cooperate. Here, *defection* (which occurs when agents renege on their agreements with others) is unlikely; indeed, agents that are unwilling to trust others may miss opportunities for cooperation because of this. Moreover, in these situations, the use of excessive regulation through strict enforcement of system or societal rules (or *norms*) may hinder agent interactions to such a degree that cooperation is ineffective.

However, in cases where agents are less likely to be benign, some form of *behaviour regulation* is needed, either through constraints imposed by organisational structure and norms (limiting what is possible for agents to do) or through careful analysis of potential cooperation partners through an analysis of *trust and reputation*. In the former case (when constraints are imposed by the organisational structure and norms), trust may be less important, since the system is heavily regulated through strict norms and enforcement. This is characteristic of the electronic institutions approach [2] in which agents do not have the possibility of violating norms. However, despite this, if agents are less willing to trust others, then the possibility for taking advantage of opportunities in terms of cooperation may be ruled out. In the latter case, (when constraints are imposed by placing less trust in agents with poor reputation), we have a prevalence of agents with malicious motivations but their effectiveness is curtailed because of the care taken in determining cooperation partners. Here, if there is little organisational structure and lax enforcement of norms, there is a high likelihood that agents may defect, and since there is little protection from societal or system regulation, the role of trust is vital. Typically, agents should place very little trust in others in these situations.

[1] King’s College London, United Kingdom, email: michael.luck@kcl.ac.uk

All of this describes the situation with real, open, distributed computational systems, in which trust, reputation, norms, etc., are relevant. Yet because we are dealing with computational systems, there is no reason why the techniques applied need mirror those applied in natural systems. In the case of trust and reputation, for example, Teacy et al. [10] describe techniques for assessing and identifying reliable interaction partners, even in the case of noisy or inaccurate data, and in particular in addressing the whitewashing problem where unreliable agents assume a new identity to avoid a bad reputation. Their approach is based exclusively on principled statistical techniques, and outperforms other models, yet is clearly not similar to techniques used by humans.

3 EXAMPLES

To illustrate some aspects here further, consider two examples. First, consider the activity of discovery, whether it be scientific discovery or data mining for patterns in data of supermarket customers. Both kinds of activity can be understood as being the same fundamental process, of seeking patterns in data, yet the ways in which they are realised must be very different. In the former case, repeatability of results, accuracy of data, etc, are primary if science is to progress, while in the latter case, it may be impossible to repeat experiments or to generate new data, and inferences must be drawn just the same, and acted upon. The difference lies in the way the various processes underpinning discovery are applied, being biased according to the motivations of the reasoning agent (the scientist is concerned with very different aspects) [4]. This is not social coordination, but the adoption of notions from psychology and social science may also be very relevant here, and in particular because the individual must be the starting point for social coordination.

Second, consider the now very prevalent operation of web services, offering the ability to build composite services out of individual components. Mainstream computing technology and no mention of agents, but this is still social coordination. We must choose services from a pool of those available, based on different factors including trust and reputation, and we must build an organisation that delivers the defined service over some period, requiring all those structures that give some assurance for organisation being effective [1], with its component parts working together in support of the broader objective. Orchestration, choreography, policies, etc., all have relevance here and are specific technical terms, but they are a clear part of social coordination and can be found, with similar or different labels, in natural systems as norms, organisational structures, etc.

Of course, this just the start. We see social coordination in all manner of systems, including those for effective electronic business that are underpinned by contracts [9] as norm-governed systems (based around web services or other technologies). Conversely, we can see these concepts in simulation systems that are designed not just to provide insight into human societies, but those that are aimed at understanding and developing effective means of encouraging or enforcing particular kinds of behaviours in different systems and societies: peer-to-peer systems provide just one immediately obvious example [8]

4 ELECTRONIC ORDER OR JUST ORDER?

Returning to the definition at the very start of this paper, that to coordinate is to "cause to function together or in proper order", we can see that all of this is concerned with bringing about order in computational systems. We might think of this as being *electronic order*, in line with the tradition of prefixing terms adopted from pre-computational usage with the ubiquitous "e" when applied in this context. However, as was argued at the start, this is to miss the point: just as with the other terms, order is applicable as much to computational systems as to natural systems. One might even argue that these notions of order are more important for computational systems since the scale and complexity of such systems demands the use of techniques for ensuring order much more than in many natural systems. In this respect, there is no different between them. E-order is just order, just as the electronic institutions mentioned above are just institutions: the metaphor of institution is as much a metaphor in real life as it is computationally. As we begin to relax over delegating control to machines, this distinction will become ever more blurred, and will (and must) disappear entirely.

The question to ask is not whether these are real manifestations of social coordination — so, not to question the validity of the approach in terms of its verisimilitude in relation to social science — but is instead whether there is value in applying the labels to the relevant techniques, and whether the techniques themselves have value in support of computational systems that can help to bring order from the chaos resulting from the masses of information and computational entities that are a very real part of modern life. Without these techniques, it is simply impossible to imagine progress.

ACKNOWLEDGEMENTS

Thanks to Elizabeth Black for allowing me to sound out the ideas in this paper.

REFERENCES

[1] L. Barakat, S. Miles, and M. Luck, 'Efficient correlation-aware service selection', in *Proceedings of the Tenth International Conference on Web Services*, (2012).

[2] M. d?Inverno, M. Luck, P. Noriega, J. Rodriguez-Aguilar, and C. Sierra, 'Communicating open systems', *Artificial Intelligence*, **186**, 38–94, (2012).

[3] T. Kuhn, *The Structure of Scientific Revolutions*, The University of Chicago Press, 1962.

[4] M. Luck, 'Evaluating evidence for motivated discovery', in *Progress in Artificial Intelligence, EPIA '93*, volume 727 of *Lecture Notes in Artificial Intelligence*, pp. 324–339. Springer, (1993).

[5] M. Luck, L. Barakat, J. Keppens, S. Mahmoud, S. Miles, N. Oren, M. Shaw, and A. Taweel, 'Flexible behaviour regulation in agent based systems', in *Collaborative Agents — Research and Development*, volume 6066 of *Lecture Notes in Computer Science*, pp. 99–113, (2011).

[6] M. Luck and M. d'Inverno, 'Motivated behaviour for goal adoption', in *Multi-Agent Systems: Theories, Languages and Applications — Proceedings of the Fourth Australian Workshop on Distributed Artificial Intelligence*, eds., C. Zhang and D. Lukose, volume 1544 of *Lecture Notes in Artificial Intelligence*, pp. 58–73. Springer, (1998).

[7] M. Luck, S. Munroe, F. Lopez y Lopez, and R. Ashri, 'Trust and norms for interaction', in *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics*, pp. 1944–1949. IEEE, (2004).

[8] S. Mahmoud, N. Griffiths, J. Keppens, and M. Luck, 'Establishing norms for network topologies', in *Coordination, Organizations, Institutions, and Norms in Agent Systems VII*, eds., S. Cranefield, J. Vazquez-Salceda, B. van Riemsdijk, and P. Noriega, volume 7254 of *Lecture Notes in Computer Science*, pp. 203–220. Springer, (2012).

[9] F. Meneguzzi, S. Modgil, N. Oren, S. Miles, M. Luck, and N. Faci, 'Applying electronic contracting to the aerospace aftercare domain', *Engineering Applications of Artificial Intelligence*, **25**(7), 1471–1487, (2012).

[10] W. T. L Teacy, M. Luck, A. Rogers, and N. R. Jennings, 'An efficient and versatile approach to trust and reputation using hierarchical bayesian modelling', *Artificial Intelligence*, **193**, 149–185, (2012).

Factors in the emergence and sustainability of self-regulation

Chris Bevan [1] and **Lia Emanuel** [1] and **Julian Padget** [2] and
Juani Swart [3] and **John Powell** [4] and **Shadi Basurra** [2]
{c.r.bevan,l.emanuel,j.a.padget,j.a.swart,s.s.a.basurra}@bath.ac.uk
john.powell@usb.ac.za

Abstract. We are interested in organizations whose goals do not primarily involve profit, if it even figures at all, but which instead seek to create social capital in a wide variety of forms. Such organizations have widely varying lifetimes, but without an equivalent to accountancy to analyse their state of health and their evolution, it can be hard to establish what brings them about, sustains them or leads to their dissolution.

We report on some preliminary work on the analysis of three such organizations, using three different approaches. Our aim is to see what common factors can be observed, in order to establish the basis for a normative model of organizations, that may then form the core of an agent-based simulation, through which we might explore the sensitivity to and dependencies between the factors.

1 INTRODUCTION

The aim of this project is to develop an understanding of self-regulation from the perspective of the individual and of the community in which s/he participates, and of the extent to which such regulation is or could be mediated by information technology in order to develop, sustain and enhance (digital and physical) communities.

Self-regulation guides our behaviour [4] and the manner with which we interact within a community, such as following shared values, implicit and explicit social norms and behaving in a way that is held to be socially acceptable within the community [3]. Ostrom [8] has shown in great detail how such mechanisms emerge and are sustained in physical resource-constrained communities. On-line communities too are starting to appear (e.g., slashdot, the bazaar model, wikipedia) with similar characteristics. As people continue to move more towards interacting within and integrating virtual communities into their daily lives [5], we believe there is a crucial need to establish an understanding of the self-regulatory properties of communities that can straddle the physical/digital divide e.g.,[1] rather than being constrained to operate in one or the other, as well as the benefits that might arise therefrom. Thus, we propose to examine how – and which – regulation principles and knowledge of self-organising groups translate from physical to digital, and vice versa, and which may not. In order to gain a richer understanding of community mechanisms we examine three unique communities which were identified as fulfilling the criteria of being self-regulating, as outlined below, and those which utilised varied physical and cyber interaction techniques. In this paper we present the preliminary analyses of the common organizational factors which emerged from our selected communities employing a cross-disciplinary approach. Specifically, we use a systems-based knowledge mapping (SBKM) technique [6] that identifies knowledge types underpinning self-regulating systems in the first case study (Stellenbosch Transition Group). A second approach, applied to the Liftshare scheme, employs survey techniques which aim to provide a preliminary measure of key predictors for individual and group self-regulatory behaviour patterns within the community, such as social roles and attitudes, motivational processes and interpersonal and group processes. The third case study (re-use groups), takes the least formal approach and is the least developed at this stage, being based on discussions with moderators and some very basic statistical analysis of group activities over the last seven years (since 2005), using public data [5].All these groups operate via internet message boards, which are monitored by volunteer moderators.

Our overarching aim is to identify the regulatory mechanisms that have emerged into a normative model of self-sustaining organisations. This will form the basis for future research (see Figure 1). This should then allow for the construction of simple demonstrators with the potential to explore the impact of combinations of different regulation mechanisms and the sensitivity of the key variables. Consequently, we can examine changes in self-regulation and subsequent behaviour via simulation scenarios.

The rest of the paper is organized as follows: in the next section we discuss background and related work; this is followed in Section 3 with a presentation of each case study, the methodology applied, the preliminary analysis arising therefrom and a short discussion summarising the observations in each case. We finish by drawing together the threads of the three case studies, followed by outlining direction for future work in Section 4.

2 BACKGROUND AND RELATED WORK

Self-organising systems are a unique form of social coordination. Such communities are driven primarily by the individuals within, interacting in a way which drives the community towards a shared goal or interest [6]. It is through self-regulation that individuals modulate, modify and monitor their behaviour to attain a given goal [4]. Thus, the self-regulatory processes at the individual and group level should

[1] Department of Psychology, University of Bath
[2] Department of Computer Science, University of Bath
[3] School of Management, University of Bath
[4] Business School, University of Stellenbosch

[5] A more in-depth analysis is in progress, as a result of gaining access to more comprehensive data.

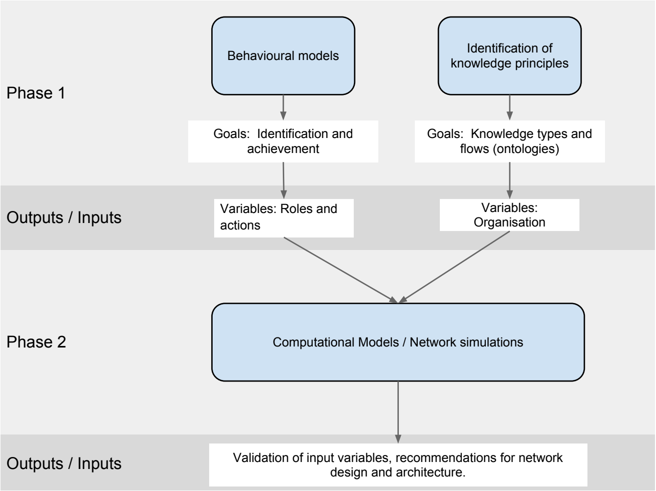

Phase 1
Behavioural models
Identification of knowledge principles
Goals: Identification and achievement
Goals: Knowledge types and flows (ontologies)
Outputs / Inputs
Variables: Roles and actions
Variables: Organisation
Phase 2
Computational Models / Network simulations
Outputs / Inputs
Validation of input variables, recommendations for network design and architecture.

Figure 1. Project outline

dictate the dynamics of a self-organising system. There are several factors which characterise the process of self-regulation and the dynamics of a self-organising system. We drew upon these principles to define a self-regulating community.

Three primary factors of successful self-regulation are *goal setting*, *self-monitoring* and *motivation*. Goal setting refers to identifying a defined goal which initiates self-regulation, and in doing so initiates action to attain that goal [1]. Within a community context, the group must share a common goal(s). Self-monitoring acts as a feedback system, allowing an individual to monitor when it is necessary to adjust an action or behaviour in order to attain ones set goal [8]. At a group level, there must be some type of environment or means to monitor progress on set goals. Lastly, an individual must have the motivation to attain a set goal, without which can lead to the failure to regulate behaviour [3]. Behncke [4] suggests lack of motivation can include trying to attain a non-realistic goal (or one you believe you have already achieved), as well as lack of incentives. Thus, the majority within a self-regulating group should have a motivational investment in the groups shared goals. Whilst there are numerous ways in which failure to self-regulate successfully may occur [2], many essentially equate to either (i) a failure to recognise action is necessary to address a need or a goal, or (ii) an inability to modify / continue appropriate action to attain a goal. It is this recognition of a need or goal by many individuals that brings together self-organising systems for collective action.

2.1 Self-organizing systems

Higher level principles of self-organising systems are also, unsurprisingly, centred around the individuals that make up a system. Lucas [6] defines two factors of self-organising systems that are particularly pertinent to human-based communities; autonomy and importance of interaction. Self-organising groups typical grow in a horizontal, not hierarchical structure [10]. Individuals are brought together through shared ideas or goals, independent of external organisations. As such, a self-regulating group cannot be formed or grow on the basis of the members fulfilling a requirement imposed by a hierarchical system. The group must instead form more organically, out of a shared need or goal. Similarly, perpetuation or growth of the group would conceivably occur through initiating new goals and/or adapting current goal to continue active participation within the group. The quality of the interaction between individuals within the group is also key for the system to function. Namely, the attainment of goals or outputs in the group cannot be dependent on one person or a small number of individuals actions [6], the group as a whole is needed to achieve a common goal. Otherwise, there is no need for the existence of the group. This highlights the question of how to ascertain what is that small number – or critical mass – by which the vitality of a group might be measured and the chances of it achieving its goals assessed.

2.2 Criteria for self-regulating communities

Using the above principles, we have set out the following criteria as a definition of a self-regulating community:

SRCC1 The group must share a common goal(s) and have the ability to communicate those goals to set them in place.

SRCC2 The group must have the means to monitor progress on set goals. This includes: (i) an effective way to evaluate whether the group is on course in meeting their aims, (ii) the means to communicate when behavioural changes are necessary to obtain set goals, and (iii) to have the knowledge required to choose alternative paths to achieve the goal.

SRCC3 Motivation to attain group goals, for instance, having some incentive in place for being a member of the group and to believe they (as a group) have the tools / ability to achieve set goals.

SRCC4 The group is autonomous. For instance, an individuals membership or involvement in that group is not due to pressure from or in obligation to a manager, institution, funding body, etc.

SRCC5 The attainment of goals or outputs in the group cannot be dependent on one person or a small number of individuals actions.

Using these criteria we selected three unique pre-existing communities as case studies to better understand the common factors in the development and functioning of self-regulating communities.

2.3 Criteria for self-organizing institutions

As noted in the introduction, Ostrom [8] has explored in depth the properties that lead to and sustain human institutions governing physical resource extraction. At this point, it is an open question whether the criteria above and the scenarios below can be captured by an adaptation of Ostrom's principles, which would essentially depend on whether the scenarios can be expressed in terms of resources and whether those resources, centred as they are around social capital and values, have instrinsic properties that make them similar or different from physical resources.

A timely analysis of Ostrom's principles (see Figure 2) is presented in [9], with the aim of showing, through simulation, that they are necessary and sufficient for efficient resource allocation and the sustainability of a single institution, because the individuals and the institution are co-dependent. In particular, if all agents' behaviour is compliant, then principles 1–3 are sufficient, but if not, principles 4–6 are also required to act as a brake on behaviour that could lead to the collapse of the resource and thereby the institution. Principle 7 is

OP1 Clearly defined boundaries: Those who have rights or entitlement to appropriate resources from the CPR are clearly defined, as are its boundaries.
OP2 Congruence between appropriation and provision rules and the state of the prevailing local environment.
OP3 Collective-choice arrangements: In particular, those affected by the operational rules partici- pate in the selection and modification of those rules.
OP4 Monitoring, of both state conditions and appropriator behavior, is by appointed agencies, who are either accountable to the resource appropriators or are appropriators themselves.
OP5 A flexible scale of graduated sanctions for resource appropriators who violate communal rules.
OP6 Access to fast, cheap conflict-resolution mechanisms.
OP7 Existence of and control over their own institutions is not challenged by external authorities.
OP8 Systems of systems: Layered or encapsulated CPRs, with local CPRs at the base level.

Figure 2. Ostrom's Principles for Enduring Institutions (from [9])

taken for granted, while principle 8, which addresses multiple institutions and the relations between then, is left for future work.

While there is no identifiable Common Pool Resource (CPR) – or at least, not yet – in our scenarios, there is a plausible correspondence between the values *behind* Ostrom's principles 1–4, 7 and 8 and the criteria set out in section 2.2, although they are structured relatively differently, having been established independently.

The ability to map the established principles set forth by Ostrom regarding the sustainability of institutions onto our own criteria of self-regulating communities, based on principles of self-and group-regulation, will provide additional insight into both the current understanding of self-regulating communities as well as highlight gaps in our knowledge around new regulatory mechanisms. This is particularly pertinent as these types of institutes/communities shift more and more into a virtual environment. Thus, the possible connections and insight arising from the communities we selected based on our own self-regulating community criteria to that of Ostrom's principles are the subject of current and future work.

3 CASE STUDIES

3.1 The Stellenbosch Transition Group

The Transformation Group of the University of Stellenbosch Business School (USB) was established in 2011, in response to a collective feeling that issues of diversity were under-reported. The group originated from a discussion with the Director of USB about the suppression of conversation about transformation, the ambition to improve the balance of white vs. South African persons of Black, Coloured and Indian ethnicity (BSI) posts and general openness about conversations around this process. After considerable debate a surprising conclusion was reached, namely that the group should not, in fact, be driven from the top of the organisation, but that it would gain more credibility in its conversations if it were seen to emerge from the 'body of the church'. In other words, to be an autonomic entity, which was self-governing and self-establishing. The group started with two members, and provided a discussion events of a round-table nature to which a minority of new informants/members were added from within USB, differing at each subsequent round table. This medium of discussion has enabled the replication, almost franchising of the transformation conversation. Set in respect of its values and even vocabulary by the original two members has meant that there has been a consistent framing of the transformation agenda. This diffusion of ideology seems to be central to the continuing self-identity of the group.

3.1.1 Methodology and Preliminary Results

The group was invited in late August 2012, to take part in an in focus group based exploration of their work and organisation using a System Dynamics based approach used extensively for strategic definition (and particularly action identification and for knowledge mapping) in organisations. Called Systems-based Knowledge Management (SBKM), it is a straightforward process of identifying causal links in the operation of, in this case, the Transformation Group, so that a model is built up, set by step, of the way in which the group operates (in this case how the transformation conversation becomes more open). There were six members present, including the two original members. This group constituted the most active, central core of the group's participants.

The results of applying the SKBM approach are shown in Figure 3, where solid links (blue and green) denote positive influences and dotted (red) are negative. The blue and red links were identifed by the group in discussion, while the green links were added as a result of post-hoc analysis and subsequently confirmed by the group and in effect have the same status as the blue links. The figure present here follows the deletion of nodes without inward and outward links, since these cannot contribute to the closed cycles of causality, whose discovery is the point of the exercise. Many of the loops present in the model are effectively duplicates, but there are 12 distinct and significant loops present, which together lead to the identification of seven key knowledge types, seen as the properties needed for this group to function and self-organise. Specifically, knowledge of:

KT1 **Qualities of autonomy:** Knowledge and recognition of the need to operate outside the structure, processes and politics of formal control.
KT2 **Energy, voice and continued freedom:** Knowing how to generate momentum within the membership via open and emotional conversation (where participants can disagree).
KT3 **Creation of coherence:** Knowing how to create coherence as a result of open, participative conversation (where agreement is reached)
KT4 **Growth dynamics:** Knowing how to balance size and inclusivity/growth.
KT5 **Continuity – Importance of linking past and future:** Knowing how to establish initiation processes wherein which each member shares their own resources and feels a sense of ownership/belonging to the group.
KT6 **Clarity of purpose:** Knowing how to structure a clear action agenda which facilitates momentum and growth.
KT7 **Heterogeneity and homogeneity:** Knowing when and how to increase heterogeneity in the group in order to secure growth.

A complete presentation of the loops and the knowledge types would take more space that is presently available, so we focus on the extraction of the loops supporting KT2 as an example (Figure 4).

Examining loop 2, in detail, we can see it contains three integrated loops with common elements. All are concerned with the dynamics

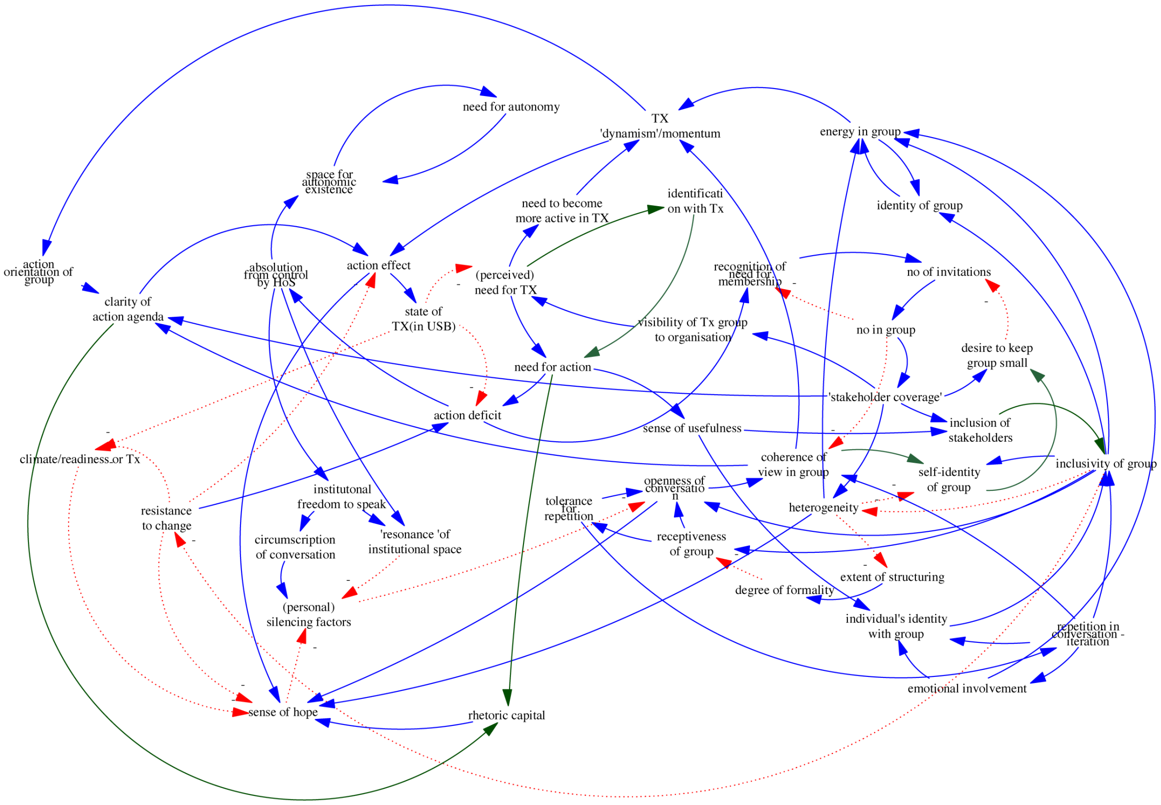
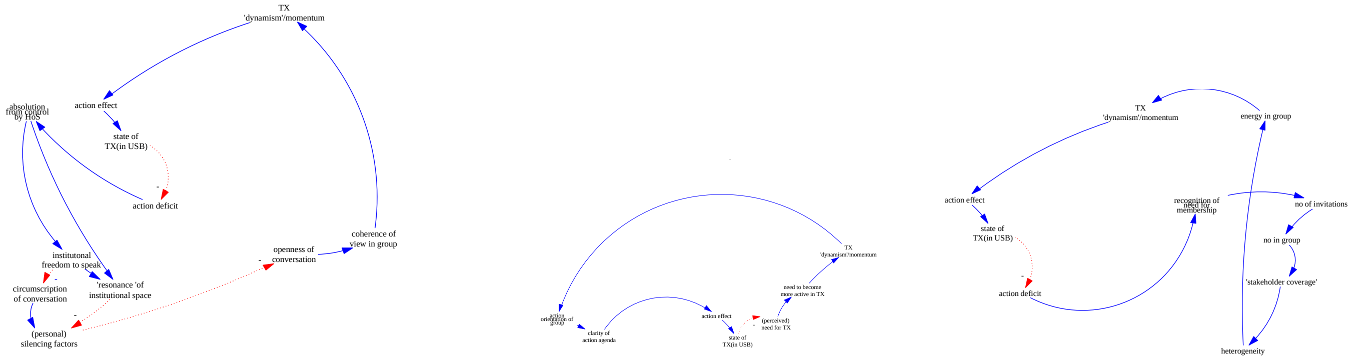


Figure 3. Result of the Systems-Based Knowledge Mapping process



(a)

(b)

(c)

Figure 4. Loops 2, 3, 4 (a) 11 (b), and 12 (c), corresponding to KT2

of the ‘momentum’ of the group, that is, its sense of forward movement and success. As the group achieves momentum (another word used was ‘traction’) this has an effect, in that the conversation which it seeks to engender improves in spread and richness. This (says the group) then improves the state of transformation in USB, primarily because the surfacing of transformation issues itself improves the way in which previously disadvantaged colleagues are treated. It is a tenet of transformation studies that making the privileged aware of the coercive nature of their privilege is itself a step towards avoiding that coercion.

This inherently-owned action effect then reduces the need for action, (action deficit). Interestingly, the absolution of control by the Head of School, deriving as it does from the need for action, is thereby reduced (i.e. as the need for action is reduced, the need for autonomy of the Transformation Group also reduces). Counter-intuitive as this is, it can be observed in the level of autonomy of the group as the transformation conversation becomes freer in the School.

The loop then divides into three paths. Loop 02 passes through institutional freedom to speak, circumscription of conversation to personal silencing factors. What is being tacitly observed by the group is that as the autonomy of the group changes (in the sense of its freedom from Head of School influence), its ability to erode the circumscription of conversation alters; a more autonomous group sees less circumscription and reduces the personal silencing factors in the School.

3.2 Liftshare

Liftshare.com is a community that straddles the physical/digital divide in that members interact in both on-line and off-line environments. Established in 1997, Liftshare currently has over 350,000 registered members. The Liftshare network enables individuals to find other people in their area to car-share, (either as a driver or as a passenger) using on-line messaging to coordinate the process. Once individuals find a car-share partner(s), they then meet in the physical world and travel together to a shared destination. Thus, the group's primary common goal is to organise and complete a shared journey successfully with other Liftshare users. The community is completely self-sustained by the members of Liftshare, and as such presents an interesting self- and group-regulation dynamic that meets our criteria for self-regulating communities.

3.2.1 Methodology and Preliminary Results

Twenty-four liftshare users (9 males, 15 females; age M = 32.08, SD = 10.02) completed a survey that was comprised of four discrete sections specifically regarding the respondents' involvement with Liftshare. In the survey, group members were asked about: (i) the role they play in the community and their actions toward achievement of a successful liftshare, (ii) goal monitoring and goal achievement, and (iii) self- and group related- regulation processes. Self- and group-regulation was measured using three psychometric scales: the Bridging Social Capital scale [11], which indicated the extent to which respondents feel the liftshare community promotes contact with a broad range of people, view themselves as a part of the broader group and diffuse reciprocity within the community. Perceived Organisational Support and Reciprocation Wariness scales [7] assesses the extent to which liftshare users perceive that the community values their contributions and cares about their well-being, and the extent to which users may be hesitant to accept or extend help as well as concerns over exploitation, respectively.

Role and actions: a Liftshare user can take one of three roles: (i) they can seek lifts from others, (ii) offer lifts to others, or (iii) both seek and offer lifts. Each role offers a different commodity to the community. Those seeking lifts do not have a car to offer as a resource to the group, but they are expected to help their fellow liftsharer pay for petrol. Conversely, those that offer lifts do not rely on the community to get to a destination (as they have a car), but the community improves their travel experience. As one user stated, "*It [Liftshare] has saved me a fortune and introduced me to some great people*". Those that both seek and offer lifts can be seen as a more versatile member of the group and potentially benefitting the most out of being a member of the community. They are able both to offer the resource of a car and are willing to share a journey with another resource-holding member. Examination of the mean ages of Liftshare users by role using analysis of variance also revealed that those in the role of both offering and seeking lifts are marginally older (M = 36.2 years) than their lifts offering (M = 35.2 years) and lift seeking (M = 25.8 years) counterparts, F (2,18) = 3.48, p = .05.

We explored the actions that members take towards the achievement of a shared journey by examining the scope of interaction that they have with the community. This included the number of journeys that they made in the last 6 months, and the number of travel partners that they typically interact with. Preliminary results showed that the role a member plays did not statistically differ in terms of the scope of interaction they have with the community. However, heavy users of Liftshare (e.g. 15+ journeys made) tend to travel with the same person, or same 2-3 people within the community, whereas less frequent users tend to have a higher number of different travel partners, r = -0.53, $p < .01$. This may suggest the formation of pockets within the community around those who interact with the Liftshare community more frequently. However, this may pose a problem for growth dynamics within the community, as one member stated, "*I havent found it [liftshare] that useful as most people I contacted were already in a liftshare and werent looking for anyone else*".

Goal monitoring and goal achievement: in order to organise and complete a shared journey, effective communication is needed to monitor that goal. The Liftshare community utilises an online messaging system that allows users to post journeys they will be making as well as the role they play in that journey (seek, offer or both seek and offer lifts). Liftshare members can then search all journeys posted within the community and contact other individuals via private message to arrange a liftshare. Survey respondents rated the effectiveness of this system as 'average' overall. However, 75% stated they had never experienced a miscommunication or missed journey once a liftshare had been agreed upon. Furthermore, members that rated the messaging system as being most effective were related to reporting that Liftshare had substantially improved their travel or commute (r = 0.47, p = .02). This suggests the ability to monitor the organisation of a liftshare journey through effective communication may lead to a positive experience of Liftshare.com and achieving a member's primary goal of successfully sharing a journey. In addition, there was a trend indicating heavy users of Liftshare (15+ journeys made in the last 6 months), and those who travel with the same person or same 2-3 people in the community reported the greatest belief of achieving the goal of improved travel through their being a member of the Liftshare community (F(2,21) = 3.08, p = .07 and r = 0.49, p = .08, respectively).

The community also has five secondary goals which are made prominent on their website (liftshare.com). Each are related to successful journey sharing: saving money, having company, travel convenience, reducing pollution and improving traffic congestion. Survey respondents were asked to rank these goals from 1 (most important), to 5 (least important), as they relate to them as a member of Liftshare. Analysis showed a significant linear effect, F (1, 21) = 28.77, $p < .01$, suggesting that saving money was ranked as being the most important (M = 1.50) to members, significantly differing in importance to reducing pollution (M = 2.92), convenience (M = 3.17), improving traffic congestion (M = 3.25), and having company (M = 3.33). The ranking of goals did not differ by the members role in the community, F (2, 21) = 1.05, p = .36 (n.s). Notably, the most important goal (saving money) is a relatively individualistic goal or incentive for being a part of the Liftshare community, whereas the second most important goal (reducing pollution) is collectivist in nature. This may suggest a self-organising community needs a variety of goal incentives, both personal and communal, to motivate the majority of the group population. This possibility is further supported in our preliminary results in examining differences in self- and group-regulation processes.

Self- and group-regulation processes: individual differences in bridging social capital – an indicator of members' feeling that the Liftshare community promotes interaction with diverse people, a sense of community and diffuse reciprocity – suggested that different goal incentives vary in importance for different members. Those members who felt more strongly about the importance of social capital tended to rank saving money as a less important goal (r = .38, *p* = .06), instead tending to rank having company on a journey (r = -0.51, *p* = .01), and the convenience of sharing a lift (r = -0.39, *p* = .05) as being more important goal incentives of being a member of Liftshare. In addition, members that reported feeling greater organisational support from the Liftshare community, such that they perceived the community valued their contributions and cared about their well-being, also tended to believe that their involvement in Liftshare improved their travel/commute (r = 0.49, *p* = .02).

Respondents reported relatively low levels of feeling apprehensive / being uncomfortable (M = 2.33) about sharing a journey with someone they met through Liftshare.com (1-5 scale, 1 indicating low levels and 5 indicating high levels). However, those who reported higher levels of apprehension/discomfort journey tended to have higher levels or reciprocity wariness, (r = 0.54, *p* = .01). Thus, members in the community that are generally hesitant to accept or extend help maybe less comfortable in the actions necessary to attain this community's common goal.

3.2.2 Discussion

In summary, several themes have emerged so far regarding roles, actions and regulatory processes within the Liftshare self-regulating community:

1. The role an individual fulfils is dictated by the different resources that they provide to the community (e.g. a car, helping to pay travel costs etc). Higher value resources (e.g. a car) tended to be provided by older members of the community.
2. Members who frequently interact with other members in the community tend to form smaller group links (e.g. always sharing a journey with the same person(s). Members who interact less frequently with the community tend to come in contact with a broader spectrum of other community members. There are potential issues here for growth of the community.
3. The current Liftshare communication system of journeys available and private messaging has room for improvement. Members who were able to efficiently communicate perceived themselves to be more successful in their ability to travel.
4. Both individualist and collectivist goals may be necessary incentives to motivate a diverse community.
5. Differences in member need for social capital and community support may influence the importance of goals, incentives, and actions put in place to achieve the communitys common goal(s).

3.3 Freecycle/Freegle: re-use groups

There are numerous local and internet-based groups that exist to try to encourage re-use in place of sending items to landfill. We focus here on Freecycle [6] and Freegle [7]. Freecycle started in Arizona in 2003 and established itself in the UK in the same year. The UK activity has since split, with about 60% of UK groups now operating under Freegle, a UK registered charity, and the remainder being administrated by the international Freecycle organization.

3.3.1 Methodology and Preliminary Results

We noted in the introduction that the approach taken to the examination of this group is less principled and less scientific in what has taken place to date. The primary sources of data have been the Freecycle and Freegle websites from which data about country presence, number of groups, group sizes and message volumes hae been taken. These metrics form the basis for a preliminary analysis of the vitality of a group. There is also anecdotal evidence from group moderators regarding the creation of new groups. There are some role similarities with the liftshare scenario, in that individuals can: 1. seek goods 2. offer goods 3. seek and offer goods. It would appear that in practice, many people are sinks or sources of goods, but that relatively fewer are both. As with Liftshare, there are both individual and collective incentives: to save money and to reduce landfill. Other factors, no doubt, also play a part, but need surveys for appropriate identification.

Out of the 370+ UK groups ($\approx$1.5M members), we have selected 10 groups at random that started in 2005 and that have a current membership of more than 10,000. As can be seen from the difference between total messages and average messages (Figure 5), size is not always correlated with activity. What is also interesting is that activity seems to peak in 2008-2009 and has been declining, but lately quite slowly, since then. Since all we have is message counts, we can only hypothesize about the reasons behind this fall. One possibility is that activity tracks, with some lag, the state of the economy. It might be expected that freecycling might increase in an economic downturn, but although more people will seek goods in lieu of paying for them, at the same time fewer people will offer goods – making do with what they have rather than replacing. In consequence, overall message counts drop. There may also be a technological explanation: (at present) the only data we can obtain relates to Yahoo-hosted Freegle groups, but the last two years have seen the migration to Freegle's own hosting service and the addition of two new channels in the form of Facebook and Twitter. Anecdotal evidence is that membership increases are observed when these channels are added to a given group and that message volume also rises[8].

3.3.2 Discussion

It was noted above that internet mediated re-cycling of unwanted goods started in 2003. Although there are claimed to be $\approx$5,000 groups worldwide across $\approx$100 countries, the main metrics (age, size, activity) are highest in the US, the UK and other Anglo-Saxon countries (NZ, AU). Indeed, groups in other countries appear often to be centred around Anglo-Saxon communities. Thus we posit another factor that may influence the sustainability of a self-regulating community: the cultural situation – or at least, how the goals of the group and the associated incentives align or not with the cultural values of its situation.

[6] `http://www.freecycle.org`, retrieved 20130209.

[7] `http://www.freegle.org.uk/`, retrieved 20130209

[8] We are currently seeking access to the data for these channels in order to extend the analysis.

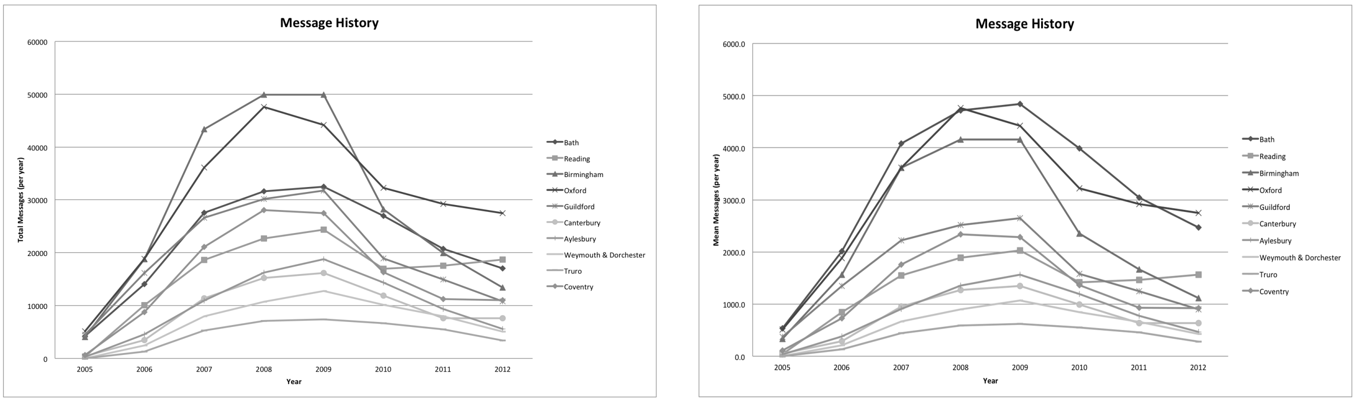


Figure 5. Total and average message counts for a selection of UK Freegle groups with a current (2013) membership $> 10K$

4 DISCUSSION AND FUTURE WORK

In line with our aim of achieving a better understanding of the common factors in the development and functioning of self-regulating communities across the physical/digital divide, the three reported case studies yielded five primary regulatory mechanisms from which we can begin to move forward in outlining a normative model of self-sustaining organisations. We summarize our observations about each of these as follows:

1. Firstly, there are distinct roles for members, which do have the ability to overlap (e.g. seek and offer lifts or goods). The current preliminary data suggests available resources of individual members may at least partially drive the role they take in the community. This factor is in line with SBKM finding which suggest knowledge of when and how to increase heterogeneity can secure growth. It remains unclear, however, what forces may trigger the need to increase heterogeneity within roles. Further research is needed to address both this and to identify the critical mass not only for the group as a whole, but fulfilment of distinct vs. overlapping roles in a community's ability to sustain and grow.
2. Second, the ability to balance the size and inclusivity/growth of a community (KT4) must be considered. Our data suggests members who frequently interact in the community may start to form smaller, exclusive, groups which in term may result in decreased growth.
3. Third, members who perceive they have the ability to communicate adequately with other group members appear to believe the community facilitates the achievement of a common goal. The ability for communication channels to evolve and adapt seem to increase community vitality. This is in line with the SBKM findings, that tools or people within the group much be able to create coherence as a result of open, participative conversation (KT3).
4. Fourth, members need both individual and collective incentives to maintain activity/vitality of the group. It is possible the individualist incentives facilitate early action from the members, as these tend to be achieved in the short term (e.g. save money), whereas collectivist incentives may act as a long term motivators for more abstract goals (e.g. reducing landfill waste, reducing pollution). These longer term motivators will hold different weight with different individuals within the group, such as those who tend to perceive value in social capital, or have different needs in their sense of ownership to the group (KT5).
5. Fifth, although self-regulating communities are autonomous, the influence of outside sources must be considered. Particularly in communities which straddle the physical/cyber divide, members of cyber communities typically participate in any number of cyber and physical communities. Economic climate and cultural situation needs to be considered, especially considering the possibility that individual circumstances within a given community could vary widely in these two respects. It is possible that the knowledge to structure a clear action agenda to facilitate momentum within the group (KT6) may allow the community to minimise the effects of exogenous influences.

We believe that the next steps in this line of research need to incorporate insights from these and further case studies on the mechanistic factors that allow self-regulating communities to sustain and grow. One potential avenue is to employ the currently reported themes, in tandem with Ostrom's principles and the self-organising criteria identified here into the generation of an ad-hoc / peer-to-peer networking model, where the survival and growth of the network is dependent upon the effective / sustained sharing of (computing) resources (storage, processing etc). Through simulation of such a network, we would hope to refine further the impact of the mechanisms so far identified, while also providing the means to rapidly evaluate other mechanisms that emerge from our continued research in this area.

The ability to rapidly add and evaluate self-regulatory mechanism in a simulation scenario will provide insight into a community's ability to change key principles [8] such as, boundaries, resource allocation, selection and modification of rules, and shifts in external authority challenges, without detrimental social effects on the development and sustainment of that community.

The above are only a few of the questions that need to be addressed in order to refine our understanding of self-organising systems. We believe the potential for this area of research for both social and computer scientist will incite an in-depth and lively discussion.

ACKNOWLEDGEMENTS

This work was partially supported by the Engineering and Physical Sciences Research Council (UK).

REFERENCES

[1] K. Barzilai-Nahon and S. Neumann, ‘Bounded in cyberspace: An empirical model of self-regulation in virtual communities’, in *System Sciences,HICSS. Proceedings of the 38th Annual Hawaii International Conference on*, p. 192b, (January 2005).

[2] R. F. Baumeister and T. F. Heatherton, ‘Self-regulation failure: An overview’, *Psychological Inquiry*, **7**(1), 1–15, (1996).

[3] R. F. Baumeister and K. D. Vohs, ‘Self-regulation, ego depletion, and motivation’, *Social and Personality Psychology Compass*, **1**, 115–128, (November 2007).

[4] L. Behncke, ‘Self-Regulation: A brief review’, *Online Journal of Sport Psychology*, **4**(1), (2002). Retrieved 20130206 from `http://www.athleticinsight.com/Vol4Iss1/SelfRegulation.htm`.

[5] Danah M. Boyd and Nicole B. Ellison, ‘Social network sites: Definition, history, and scholarship’, *Journal of Computer-Mediated Communication*, **13**(1), 210–230, (October 2007).

[6] C. Lucas. Self-Organising systems (SOS), 2008. Retrieved 20130206 from `http://www.calresco.org/sos/sosfaq.htm`.

[7] P. D. Lynch, R. Eisenberger, and S. Armeli, ‘Perceived organizational support: Inferior versus superior performance by wary employees’, *Journal of Applied Psychology*, **84**(4), 467–483, (1999).

[8] Elinor Ostrom, ‘Coping with tragedies of the commons’, *Annual Review of Political Science*, **2**(1), 493–535, (1999).

[9] Jeremy Pitt, Julia Schaumeier, and Alexander Artikis, ‘Axiomatization of socio-economic principles for self-organizing institutions: Concepts, experiments and challenges’, *ACM Trans. Auton. Adapt. Syst.*, **7**(4), 39:1–39:39, (December 2012).

[10] M. J. Wheatley, ‘Leadership of self-organised networks: Lessons from the war on terror’, *Performance Improvement Quarterly*, **20**(2), 59–66, (2007).

[11] D. Williams, ‘On and Off the ’Net: Scales for social capital in an online era’, *Journal of Computer-Mediated Communication*, **11**(2), 593–628, (2006).

Experimental Interaction Science

Flavio S. Correa da Silva[1] and **David. S. Robertson**[2] and **Wamberto Vasconcelos**[3]

Abstract. We characterise an emergent area of research which amalgamates different research traditions from within Computer Science and Economics. Experimental Interaction Science concerns the investigation of methods and techniques for experimental construction and analysis of models for interactive systems comprising humans and machines. We view Experimental Interaction Science as a conceptual framework used to structure a well specified research methodology, present its roots and methodological grounds, and then identify existing initiatives which can be reconstructed within this conceptual framework and, therefore, brought together within a single and coherent conceptual structure.

1 Introduction

In this article we characterise an emergent research area, coined Experimental Interaction Science, which amalgamates different research traditions from within Computer Science and Economics. Experimental Interaction Science concerns the development of methods and techniques for experimental construction and analysis of models for interactive systems comprising humans as well as machines.

In both Computer Science and Economics we identify a historical methodological research divide, which has only recently started to converge. Interestingly, the methodological convergence in both fields has occurred in subfields whose primary concern is interactions involving heterogeneous parties.

Within Computer Science, the roots of the methodological divide to which we are referring can be traced to Stanford University in the early 1960s, when Douglas Engelbart created the Augmentation Research Center at SRI and John McCarthy created the Stanford Artificial Intelligence Laboratory [29]. McCarthy proposed a machine-centric approach to technological development, based on principles of optimal performance grounded on mathematical modeling and aiming at improving (and correcting) human capabilities and behaviour following those principles, whereas Engelbart proposed a human-centric approach to develop computer technology, aiming at augmenting human capabilities through computational devices.

The research methodologies adopted by these groups were contrasting. McCarthy and colleagues (and followers) adopted a model-oriented approach, in which models were built based on principles of optimal rationality, and then used to explain and suggest improvements on the behaviour of existing systems, including human beings. Engelbart and colleagues (and followers), in turn, adopted an observation-oriented approach, in which descriptions of observed phenomena were empirically compiled, and then used to guide the synthesis of explanatory models.

Within Economics, similar historical evolution has occurred [20]. Neoclassical economics has proposed models of behaviour and interaction, based on principles of optimal rationality and idealised premises of available computational resources and perfect information, whereas behavioural economics has been based on empirical observations of human behaviour and, as a reaction to the tenets of neoclassical economics, identification of counter-examples indicating limitations of models based on notions of optimal behaviour.

In Computer Science as well as in Economics, proposals based on models of optimal behaviour have employed models to suggest improvements on observed systems, whereas proposals based on empirical observations have employed data analysis to suggest improvements on existing models.

A topic of shared interest between Computer Science and Economics is interactions among heterogeneous parties encompassing human as well as digital agents. Within Computer Science, this topic has led to the development of studies in certain branches of multiagent systems following the optimised models tradition [19], and to the development of human-computer interfaces and social informatics in the observation oriented tradition [18]; within Economics, this topic has led to the development of classical game theory and related issues (such as mechanism design and social choice theory) following the optimised models tradition [5, 15], and to behavioural game theory in the observation-oriented research tradition [6].

In recent years, proposals have been made to bring together these research traditions, within Computer Science as well as within Economics. In Computer Science, we find the Services Science Manifesto [3] and the Social Computer Manifesto [7], proposing the design and development of complex systems made of humans as well as computational devices. In Economics, we find the work of Nobel laureates Vernon Smith [20] and Elinor Ostrom [16], respectively on the adoption of rigorous experimental techniques in economical sciences and the modeling and analysis of complex institutions, including detailed research programmes to ground the obtained results and conclusions.

The proposed methodological tenets suggested by Smith and by Ostrom can be transferred to Computer Science with little adaptation. In the present article we explore this possibility. In Section 2 we discuss a little further the two research traditions in Economics referred to in the previous paragraphs, and how they have been combined in the line of research coined *Experimental Economics*. We then turn to the same issue as it has occurred in Computer Science. In Section 3 we discuss some mathematical tools which can be useful to build operational models for analysis of empirical results, introduce our suggested approach to combine the two traditions within Computer Science, and outline how this approach can be used to frame existing work within a single and coherent conceptual structure. Finally, in Section 4 we discuss some potential applications of our proposed approach and draw some conclusions.

[1] University of Sao Paulo Brazil 05508090 (fcs@ime.usp.br)
[2] University of Edinburgh UK EH8 9AB
[3] University of Aberdeen UK AB24 3UE

2 Research traditions in Economics and in Computer Science

Sci-fi aficcionados may recall the frequent arguments between Dr. McCoy and Mr. Spock in the vintage Star Trek series. In some episodes purely (and perfectly) rational behaviour as featured by Mr. Spock solved complex situations and saved the day, and in many other episodes this same behaviour (sometimes expressed through slogans and propositions typical to military doctrine) was deemed inadequate or insufficient to solve other situations, and had to be complemented by "human" traits to build proper solutions.

Interestingly, this conflict between pure logic and "human" behaviour has been recurrent in scientific circles for decades. In Economics, for example, we have neoclassical scholars and technocracy defending the engineering of economic systems based on rational agents – the *homo economicus* – capable of accessing infinite computational power, unlimited sources of information (including reflective information about one's mental states), clearly stated and measurable goals and sharply defined rules to constrain their actions [5, 15]. Complex organisations are designed based on these premises, and then the consequences of relaxing or limiting some premises are analysed – for example, the computational complexity of finding equilibria in game theoretic formulations has been extensively studied, thus relaxing the premise that infinite computational power is always available.

From this perspective, deviations from *homo economicus* are treated as defects one has to live with, and pure and perfect rationality are treated as ideals to be reached. A contrasting view has been proposed by behavioural economics [1, 2, 13], in which social behaviour is observed through controlled experiments and presented as the sole genuine human behaviour to be considered. Orthodox economists have been challenged by this view and the corresponding experimental results.

These conflicting views have led to the organisation of two research communities: one focusing on building models based on normative views of rationality and then searching for means to overcome the imperfections of real systems so that they can get asymptotically closer to the perfectly rational ones, and the other focusing on making experiments to identify actual (sometimes called "irrational") behaviour, this way suggesting that formal models of rationality would have little place in helping to understand actual human behaviour.

The way to reconcile these views has been proposed – from different perspectives – by well-know authors such as Daniel Kahneman [12], Elinor Ostrom [16] and Vernon Smith [20], all three of them Nobel laureates. The general idea has been to acknowledge human rationality as being complex and sophisticated, and possibly only accessible through experiments, instead of deeming humans as "irrational". Models of rationality should, therefore, be grounded on empirical observations, this way characterising *Experimental Economics*, and its counterpart in the design and analysis of interactions, *Behavioural Game Theory* [6].

In Computer Science, the same conflict can be identified in the methodological divide which separated Human-Computer Interaction and Artificial Intelligence, as presented in section 1. The need for convergence between the research traditions championed by Engelbart and McCarthy has been characterised [29], and recently proposals to achieve this convergence have been put forward (*e.g.*, [3, 7]). There is still room, however, to articulate a research methodology for the convergence. In the present article we start one such articulation, following the lines of *Experimental Economics*, hence the title of this article and of the next section.

Many interesting proposals have been developed providing further evidence that the convergence of the research traditions outlined here can be profitable within Computer Science [11]. Interestingly, this convergence has been less observed in the analysis of interactive heterogeneous systems as characterised here, for which the appropriation of the research methodology developed in Economics can be best and most easily performed. Early attempts – which have indeed inspired the propositions put forward in the present text – have partially combined the model- and the observation-based approaches, however with no preliminary articulation of a research programme as proposed here:

- A socio-technical system has been proposed for emergency prevention and relief when floods occur in Trentino, northern Italy [7].
- Human-robot interactions have been studied to enable robots with complex social interaction skills, such as improvised dancing [22].
- Synthetic characters have been developed in computer games to build empathic relations with human players [4].

3 Experimental Interaction Science

We focus on heterogeneous complex systems comprised by humans as well as digital agents, and study these systems based on empirical analysis of their behaviour and patterns of convergence towards equilibria in relations involving all participants. Our goal is to build operational and executable descriptive models, which can be used for scientific analysis and understanding of the systems under consideration, as well as for technological development of effective organisational systems to serve specific purposes and to enhance human relations in general. In this sense, from a methodological point of view, our proposition belongs to *Pasteur's Quadrant*, following the terminology suggested by Donald Stokes [21].

In order to build operational and executable models based on empirical data, we need the appropriate conceptual tools, capable of representing processes for data analysis and empirical model building, as well as generalisation and abstraction of concepts and the possibility to perform effective inferences on data and their abstractions. In other words, we need conceptual tools which are sufficiently expressive in order to capture and represent reasoning processes which can be identified in both research traditions as characterised in the previous paragraphs, in coherent and integrated fashion.

We have found two conceptual frameworks, whose combination can be a suitable candidate for the conceptual tools we need. Both frameworks have been developed around a little more than a decade ago, and focus on similar problems. Interestingly, these conceptual frameworks do not refer to each other, and both have had small impact on further development of theories and models for interactions based on a convergence of the research traditions referred to here. Both conceptual frameworks share the goal of integrating and articulating deductive and inductive reasoning, as well characterised by Janssen[4]:

- Deductive reasoning (which relates to models based on optimal rationality as characterised above):
  Theory → hypotheses → observation → confirmation.
- Inductive reasoning (which relates to models based on empirical analysis as characterised above):
  Observations → patterns → hypotheses → theory.

[4] Janssen, M. A. Games & Gossip (ebook), 2010, *http://www.openabm.org/book/1928/games-gossip.*

One conceptual framework is based on the notions of *Robust Logics* [23] and *Knowledge Infusion* [24]. Robust logics are built from a first order logic with a finite number of constants, a finite number of predicates and no function symbols (which, therefore, has expressive power equivalent to a propositional logic with a finite number of atomic propositions), whose semantics has been coined *PAC semantics*, or *Probably Approximately Correct semantics*. PAC semantics determines the truth-value of ground atomic formulae based on statistical sampling, in such way as to intertwine three dependent attributes:

1. An upper bound for an error rate, which characterises the ratio between disagreements between the actual truth values of formulae and a guess for these truth values based on statistical sampling, with respect to the total number of truth valuations in the language.
2. A lower bound for the probability that a certain upper bound for the error rate above can be guaranteed.
3. A lower bound for the sample size that guarantees these two probabilistic bounds, which, therefore, provides an estimate for the expected computational effort required to satisfy those bounds.

Knowledge infusion is a machine learning technique for effective synthesis of logical theories based on robust logics.

A second conceptual framework is founded on a complete reformulation of Probability Theory, replacing the usual measure theoretic foundations by game theoretic concepts. In this framework, we have probabilistic processes viewed as games against Nature, and probability measures interpreted as values worth betting by the Player (who is playing against Nature) in equilibrium strategies [17]. Game-theoretic probabilities result, therefore, from dynamic processes, instead of static set theoretic notions such as those found in measure theoretic probabilities.

Game-theoretic probabilities have been used to build *conformal prediction algorithms*, which are machine learning techniques for online prediction of future events [28].

Both conceptual frameworks focus on dynamic online processes for empirical data gathering as the basis for the construction of expressive descriptive formal theories for observed phenomena. It is interesting to observe that there is very little cross-reference between these frameworks. Their impact has also been relatively small, possibly due to their mathematical sophistication and to both frameworks having been presented through applications in rather narrow domains (word inference in computational linguistics and financial forecasting). Our perception is that they can be simplified and combined in such way as to preserve their essential attributes and capabilities, and that this combination can be used as the basis for Experimental Interaction Science.

Other conceptual frameworks can also be useful to build the foundations for Experimental Interaction Science. For example, we have:

- Recent work by J. van Benthem and colleagues, in which Game Theory as well as Probability Theory have been grounded on variations of dynamic logics, and a combination of these theories as an encompassing theory for analysis of interactive multi-party systems has been suggested [25, 26, 27].
- Recent work by J. Y. Halpern and colleagues, in which logical systems have also been proposed to ground both Game Theory and Probability Theory, and their combination has also been suggested – employing, however, logical systems which differ from those adopted by van Benthem and colleagues [8, 9, 10].

Even though a characterisation of Experimental Interaction Science as presented here is new, it has been partially formulated in documents such as the Social Computer Manifesto [7], as we mentioned previously in this article. It is also implicit in some ongoing research initiatives:

- The social computer[5] – internet-scale human problem-solving, which is an EU FP7 project focusing explicitly on the sort of systems considered here.
- Scrutable autonomous systems[6], which is an UK's EPSRC project focusing on improving the effectiveness of interactions between human users and digital agents.
- The FuturICT EU initiative[7], which is a very large and highly ambitious initiative aiming at the development of infrastructure to monitor the actual behaviour of economic agents (with special interest on macroeconomic agents and agencies), hoping to be able to build descriptive models from those observations.

Our proposition is to work out the conceptual framework as outlined in the previous paragraphs, and to make use of the internet to build controlled experiments in the form of multiplayer online social games, allowing human as well automated players, such that the behaviour of all players can be monitored descriptive models can be inferred. Examples of experiments, of increasing complexity, that can be built are:

- Strategy games designed for individual players, and matches organised in which the proportion of humans versus machine controlled (perfectly rational) players can vary between matches. Equilibrium analyses, as well as analyses of individual behaviour of players will be performed, in order to better understand interactions involving heterogeneous agents, and the effects on human behaviour of forcing competitive interactions with machine controlled agents.
- Similar experiments, based on strategy games, in which players are organised as teams; each individual player must cooperate with members of the same team and compete with members of the other teams. As above, different proportions of humans versus machine controlled players can be built, and the behaviour of individuals as well as of the system as a whole can be monitored and analysed, in order to better understand the effects on human behaviour of forcing cooperative as well as competitive interaction with machine controlled agents.
- Similar further experiments, based on games of imperfect information in which agents are allowed to bluff. As in the previous experiments, players can be human as well as automated (perfectly rational) agents. In addition to the observations considered in the previous classes of experiments, the behaviour of individuals as well as of the system as a whole can be monitored and analysed, as regards the extent to which tacit goals and ethical constraints can influence social behaviour and interactions.

Different scenarios can be considered in which the outcomes of such experiments can be useful:

1. Unmanned vehicles designed for public transportation in urban settings. In this scenario, it is likely that synthetic agents are needed to control these vehicles, whose behaviour must be carefully crafted and whose performance must be fully predictable and manageable. These agents will most likely need to interact, however, with natural agents (such as human pedestrians and drivers

[5] http://socialcomputer.eu/
[6] http://tinyurl.com/azfh762
[7] http://www.futurict.eu/

of standard vehicles), and therefore accurate models of interaction involving these different sorts of agents must be built.

2. Systems for digital entertainment and for effective dissemination of information mimicking the behaviour of human counterparts. For example, in computer games it can be interesting for a gamer to play against the computer and feel as if she is playing against a human opponent. In order to build effective synthetic agents to mimic the behaviour of human agents, at least as it refers to how synthetic and human agents relate to each other, accurate models of interaction involving human agents must be built.
3. As is well known in Game Theory [5, 6, 15, 16], interaction equilibria based on hypotheses of perfect rationality are frequently sub-optimal, and better results can be observed when interacting agents collectively build interaction protocols grounded on sophisticated models of rationality which may, however, require more complex agent modeling than perfect rationality based modeling. In order to build such models, empirical analysis of emergent interaction protocols are required.

A similar initiative to ours has been conducted by Microsoft Research [14]. Their focus, however, has been on human performance and how it contrasts with the performance of machine controlled (perfectly rational) agents. In this sense, their approach is more conservative than ours, as they align their work with the research tradition posed by behavioural economists. Nevertheless, the interesting results they have presented confirm that the approach to build experiments based on internet social games can be effective.

We believe that our approach can provide other initiatives with interesting and useful insights, which may help and inspire their own work, as we focus on the establishment of rigorous empirical methods to infer realistic models of interactions from observations. Our contributions shall come from the inferred models *per se*, which shall be useful for economists, policy makers and entrepreneurs, as well as from the characterisation of the proposed rigorous methodology which is under development, which we hope can be useful in a variety of settings.

4 Discussion and future work

We have introduced a topic for further development, coined *Experimental Interaction Science*. The proposed methodology for this topic has been strongly influenced by Experimental Economics. We have introduced some conceptual frameworks which shall be useful for the formulation of the specialised formal tools to be used in Experimental Interaction Science, as well as some concrete empirical programmes to be developed in order to further this research topic. We have also identified some ongoing initiatives which can be (at least partially) aligned with the proposed methodological views and propositions put forward in the present article.

Several technological areas can benefit from work in Experimental Interaction Science. For example:

- Group recommender systems can be developed, in which recommendations can take into account the dynamics of interactions among group members.
- Resource management in cloud computing can be enhanced, as the distribution of computational resources can take into account stochastic analyses of system behaviour from clients and the negotiation of common pool resources.
- Innovative interactive systems can be built for the personalised delivery of information to individuals, as well as for digital arts and entertainment.

Our future work shall be devoted to the detailed specification of the conceptual formal tools to be used in this area, to the design, implementation and analysis of data obtained through experiments, and to the design of prototypes for applications such as those referred to in the previous paragraphs.

ACKNOWLEDGEMENTS

The present work has been partially supported by CNPq (Brazil), and the EPSRC (UK, grant ref. EP/J012084/1)

REFERENCES

[1] D. Ariely, *Predictably irrational: the hidden forces that shape our decisions (revised and expanded edition)*, Harper, 2010.

[2] D. Ariely, *The upside of irrationality: the unexpected benefits of defying logic at work and at home*, Harper, 2010.

[3] H. Chesbrough and J. Spohrer, 'A research manifesto for services science', *Comm. ACM*, **49 (7)**, 35–40, (2006).

[4] F. S. Correa da Silva and A. F. Bressane Neto, *Affective agents for empathic interactions*, 161–172, International Conference on Entertainment Computing, Springer LNCS 6972, 2011.

[5] D. Fudenberg and J. Tirole, *Game theory*, MIT Press, 1991.

[6] H. Gintis, *The bounds of reason: game theory and the unification of behavioral sciences*, Princeton University Press, 2009.

[7] F. Giunchiglia and D. S. Robertson, 'The social computer - combining machine and human computation', *University of Trento Technical Report*, **DISI-10-036**, (2010).

[8] J. Y. Halpern, 'A computer scientist looks at game theory', *Games and Economic Behavior*, **45:1**, 114–131, (2003).

[9] J. Y. Halpern, 'A nonstandard characterization of sequential equilibrium, perfect equilibrium, and proper equilibrium', *International Journal of Game Theory*, **3838:1**, (2009).

[10] J. Y. Halpern, 'Lexicographic probability, conditional probability, and nonstandard probability', *Games and Economic Behavior*, **68:1**, 155–179, (2010).

[11] T. Hey, S. Tansley, and K. (eds.) Tolle, *The fourth paradigm: data-intensive scientific discovery*, Microsoft Research Press, 2009.

[12] D. Kahneman, *Thinking, fast and slow*, Farrar, Straus and Giroux, 2011.

[13] D. Kahneman, P. Slovic, and A. Tversky, *Judgement under uncertainty: heuristics and biases*, Cambridge University Press, 1982.

[14] P. Kohlil, Y. Bachrach, D. Stillwell, M. Kearns, R. Herbrich, and T. Graepel, *Colonel Blotto on Facebook: the effect of social relations on strategic interaction*, ACm Web Science 2012, 2012.

[15] N. Nisan, T. Roughgarden, E. Tardos, and V. V. (eds.) Vazirani, *Algorithmic game theory*, Cambridge University Press, 2007.

[16] E. Ostrom, *Governing the commons: the evolution of institutions for collective action*, Cambridge University Press, 1990.

[17] G. Shafer and V. Vovk, *Probability and finance: it's only a game!*, Wiley, 2001.

[18] B. Shneiderman and K. Plaisant, *Designing the user interface*, Pearson, 4th edn., 2005.

[19] Y. Shoham and K. Leyton-Brown, *Multiagent systems: algorithmic, game theoretic and logical foundations*, Cambridge University Press, 2008.

[20] V. L. Smith, *Rationality in Economics: constructivist and ecological forms*, Cambridge University Press, 2009.

[21] D. E. Stokes, *Pasteur's quadrant: basic science and technological innovation*, Brookings, 1997.

[22] I. S. Tholley, Q. G. Meng, and P. W. H. Chung, 'Robot dancing: what makes a dance?', *Advanced Materials Research*, **403-408**, 4901–4909, (2012).

[23] L. G. Valiant, 'Robust logics', *Artificial Intelligence*, **117 (2)**, 231–253, (2000).

[24] L. G. Valiant, *Knowledge infusion*, 1546–1551, Proceedings AAAI, 2006.

[25] J. van Benthem, 'Rational dynamics and epistemic logic in games', *International Game Theory Review*, **9:1**, 13–45, (2006).

[26] J. van Benthem, J. Gerbrandy, and B. Kooi, 'Dynamic update with probabilities', *Studia Logica*, **93:1**, 67–96, (2009).

[27] J. van Benthem, E. Pacuit, and O. Roy, 'Toward a theory of play: A logical perspective on games and interaction', *Games*, **2:1**, 52–86, (2011).

[28] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*, Springer, 2010.

[29] T. Winograd, ‘Shifting viewpoints: artificial intelligence and human-computer interaction’, *Artificial Intelligence*, **170**, 1256–1258, (2006).

Situational Analysis of Games

Corinna Elsenbroich [1]

Abstract.
Theoretical game theory has been a successful theory in economics and other social sciences. Experimental game theory, on the other hand, seems to open more problems than it solves. Almost every experimental setup results in much higher levels of cooperative behaviour than rationality allows. This paper presents an agent-based model to generate the population level outcomes of some prominent explanations of human behaviour by implementing alternatives to perfect rationality.

1 Introduction

Theoretical game theory has been a successful theory in economics and other social sciences. Experimental game theory, on the other hand, seems to open more problems than it solves. Almost every experimental setup results in much higher levels of cooperative and fair behaviour than predicted. Once we leave perfect rationality we enter the realm of psychological explanations of human decision-making and action. This paper discusses a prototype agent-based model to test population level outcomes of different psychological explanations.

2 Human Actions Reconsidered

Alternative decision procedures can be divided into those using individuals' dispositions and those using situational cues to trigger decisions and actions. Examples of dispositional explanations of decision making are motivation models [7]; [9]; [3] or approaches using collective rationality

1. Motivation Models: Preferences are amended by intrinsic motives, which may vary between individuals. Motives appealed to are for example altruism [7], fairness [9] or intrinsic aversions to inequality [3].
2. Collective Rationality: : There is a variety of collective approaches such as team reasoning [4] and [12], collective intentionality [10], [13] collective rationality [14]. What they all have in common is that they see the human as intrinsically social and collectively minded. Rather than calculating the maximal individual payoff, people reason about a situation collectively, for example, they see the collective diagonal in the prisoners dilemma payoff matrix (see Figure 2). Rather than amending the existing individual set of preferences with other motivations, collective approaches see collective preferences as the fundamental set, with individual preferences as a special case.
   Examples of situational explanations are theories using the influence of social norms on human actions, such as normative frames [1] or environmentally situational theories such as situational action theory [15].
3. Normative Frames: The idea that behaviour is not down to calculations of utilities at every decision point but that situations trigger normative frames which in turn produce behaviour patterns. The explanation of one-shot cooperation is that a normative frame of repeated interactions, where cooperation can be a winning strategy, is triggered [1].
4. Situational Action Theory: The idea that an action decision results from the interplay of environmental, situational and psychological variables. Currently a theory applied to the commission of crime only [15]. As so often in the social sciences these theory live their separate lives, each being refined, tested and promoted. Whilst some are incompatible, for most of these theories it holds that they explain certain aspects of human action.

One way to test theories in the social sciences is by the use of agent-based modelling as a virtual laboratory to generate macro effects from micro behaviours [2]. Using agent-based modelling in just such a way, these different kinds of explanations are implemented to see what the population level behaviours in game situations 'would be' if a certain procedure was indeed the underlying mechanism of human decision making.

3 Agent-based Models of Games

3.1 Simultaneous Decision Games

Using an agent-based model the population level consequences of the above explanatory hypotheses are explored, i.e. what are the population conditions in which certain kinds of cooperation/defection patterns emerge. The model implements dispositional and situational normative frames and dispositional and situational collective frames.

Rather than using payoff expectations to guide agent choices, choices are triggered by action-trees. The choice points in the action tree are either informed by an agent's attributes (dispositional) or by the agent's personal game history (situational). The agents can be more or less collectively minded and have more or less commitment to norms. If the first choice point is seen as dispositional, agents interpret a situation as collective or individual depending on their disposition alone. If it is interpreted situationally they use their personal interaction history of the past 10 interactions to inform their interpretation in addition to their disposition. The agents then decide whether a situation is norm-governed or not using the same personal interaction history or their normative disposition.

The two dispositions, collective commitment and normative commitment, are normally distributed across the population, the mean of the distributions can be varied. In the simulation, agents are paired up randomly and play a prisoner's dilemma game. Payoffs are calculated following the payoff matrix in Figure 2.

Which decision an agent makes at the choice points depends on thresholds.

[1] University of Surrey, Guildford, UK email: c.elsenbroich@surrey.ac.uk

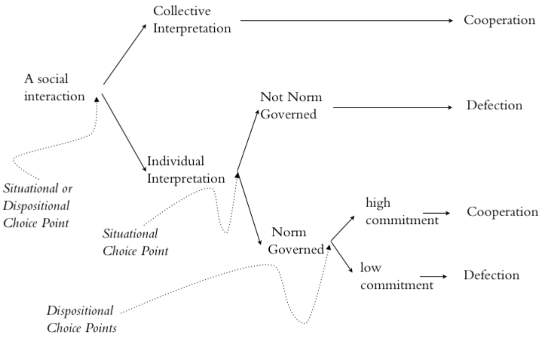


Figure 1. Prisoner's Dilemma decision tree: Whether a situation is interpreted as collective can be dispositional or situational, whether it is normative or not is situational and whether an agent uses a normative frame depends on its normative commitment.

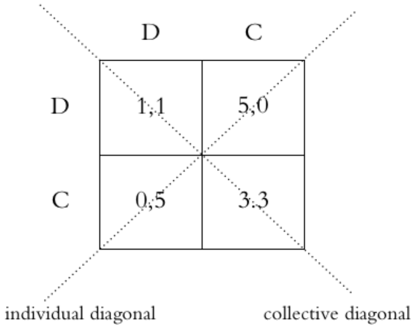


Figure 2. Prisoner's Dilemma payoff matrix.

First results running this relatively simple model show interesting interactions between the distribution of attributes in the population, the memory threshold (i.e. how many past interactions have to be cooperative to trigger collective or normative interpretations of a situation) and the different explanations. What we want to see is whether and under what conditions cooperation is established as the dominant behaviour in the population.

Cooperation can become the main choice of agents without the payoff of cooperation being higher than that of defection. Although preliminary and 'rough', this model opens an interesting space for investigating the establishment of cooperation in a population despite cooperation not paying for an individual using some explanations from the literature, namely collective rationality, normative frames and team reasoning.

3.2 Sequential Decision Games

Another highly influential Game in Game Theory is the Ultiamtum Game. In the Ultimatum Game, two players divide a windfall W of unit size 1. The players move sequentially and take on the different roles of proposer P and responder R. P offers a share of W to R. R decides whether to accept or reject W. If R accepts W is divided according to P's offer; if R rejects, neither gets anything. If agents were rational, they would a) offer a minimum amount and b) would accept a minimum amount. This is not so, as Table shows.

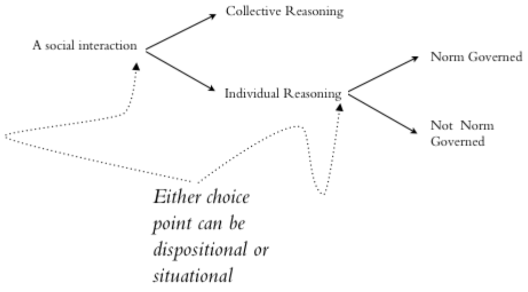


Figure 3. Ultimatum game decision tree: Whether a situation is interpreted as collective can be dispositional or situational, whether it is normative or not is situational and whether an agent uses a normative frame depends on it's normative commitment.

The two dispositions, collective commitment and normative commitment, are normally distributed across the population, the mean of the distributions can be varied. In the simulation, agents are paired up randomly and play an Ultimatum Game. Payoffs are calculated following the game tree in Figure 4.

Again, this very simple model lets us explore the population consequences of certain dispositions. Conceptually the dispositions can be linked to the individualism index and trust in Figure 5 where higher collective commitment means a lower individualism score and higher normative commitment a higher trust score in a population. Higher mean offers are a direct consequence from higher collective commitment. Mean rejections are more difficult to replicate in the model as they vary widely, and slightly erratically, between countries. In addition to the population composition consequences for offers and rejections we can see what kind of agent does well in the respective societies.

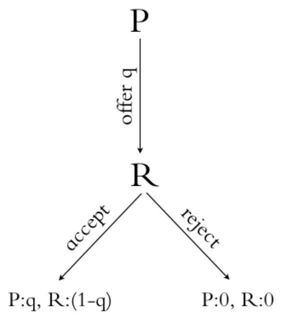


Figure 4. Ultimatum Game.

	Mean offer	Mean reject	Individualism Index	Trust
Austria	39.2	16.1	55	.32
Chile	34	6.7	23	.23
France	40.2	30.7	71	.23
Germany	36.7	9.5	67	.38
Japan	44.7	19.3	46	.42
Yugoslavia	44.3	26.7	27	.3
Netherlands	42.2	9.3	80	.56
Peru	26	4.8	16	.05
Spain	26.6	29.2	51	.34
Sweden	35.2	18.2	71	.66
UK	34.3	23.4	89	.44
US East	40.5	17.2	91	.5
US West	42.6	9.4	91	.5

Figure 5. Taken from [8, p. 177]. The individualism index is from [6] country profiles and Trust is the percentage in a country's population saying most people can be trusted (World Values Survey).

4 Conclusions and Future Work

These two simple implementations already give us some insights about the influence of some of the existing theories in the literature. The immediate future work is to scrutinize whether the operationalisations made for the implementation into ABM are faithful to the original theories, and if not, how to improve them. Also, theories such as team reasoning and collective rationality need to be properly distinguished and all the intricate differences probed as to whether they are salient or not for the model.

Future work is to develop a replicator dynamic on this model to see what agent types do well under what circumstances and what kinds of 'societies' evolve, embedding this work with the direct and indirect evolutionary accounts of cooperation (for the former see [11], for the latter [5]) and see whether and how collective and normative dispositions can evolve.

REFERENCES

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

[2] Corinna Elsenbroich, 'Explanation in agent-based modelling: Functions, causality or mechanisms?', *Journal of Artificial Societies and Social Simulation*, **15**(3), 1, (2012).

[3] Armin Falk, Ernst Fehr, and Urs Fischbacher, 'On the nature of fair behaviour', *Economic Inquiry*, **41**(1), 20–26, (2003).

[4] *Michael Bacharach: Beyond Individual Choice: Teams and Frames in Game Theory*, eds., Natalie Gold and Robert Sugden, Princeton, NJ: Princeton University Press, 2006.

[5] Werner Güth and Hartmut Kliemt, 'The indirect evolutionary approach: Bridging the gap between rationality and adaptation', *Rationality and Society*, **10**, 377–399, (1998).

[6] Geert Hofstede, *Cultures and Organizations: Software of the Mind*, New York: McGraw-Hill, 1991.

[7] David K. Levine, 'Modeling altruism and spitefulness in experiments', *Review of Economic Dynamics*, **1**, 593–622, (1998).

[8] Hessel Oosterbeek, Randolph Sloof, and Gijs ven de Kuilen, 'Cultural differences in ultimatum game experiments: Evidence from a meta-analysis', *Experimental Economics*, **7**, 171–188, (2004).

[9] M. Rabin, 'Incorporating fairness into game theory and economics', *Amer. Econ. Rev.*, **83**(5), 1281–1302, (1993).

[10] John R. Searle, *The Construction of Social Reality*, Penguin, 1995.

[11] Brian Skyrms, *Evolution of the social contract*, Cambridge University Press, 1996.

[12] Robert Sugden, 'The logic of team reasoning', *Philosophical Explorations*, **6**, 165–181, (2003).

[13] Michael Tomasello, M. Carpenter, and U. Lizskowski, 'A new look at infant pointing', *Child Development*, **78**, 705–722, (2007).

[14] Paul Weirich, *Collective Rationalty: Equilibrium in Cooperative Games*, Oxford University Press, 2010.

[15] Per-Olof H. Wikström, 'Explaining crime as moral actions', in *Handbook of the Sociology of Morality*, eds., S. Hitlin and S. Vaisey, Handbooks of Sociology and Social Research, Springer Sceince and Business Media, (2010).

Theory in social simulation: Status-Power theory, national culture and emergence of the glass ceiling

Gert Jan Hofstede[1]

Abstract. This is a conceptual exploration of the work of some eminent social scientists thought to be amenable to agent-based modelling of social reality. Kemper's status-power theory and Hofstede's dimensions of national culture are the central theories. The article reviews empirical work on the case of playing children, with a focus on field studies on pre-puberty children in several countries. The idea is to investigate emergence of glass ceiling phenomena for girls among these children. A prototype playground simulation gives a proof of concept. The conclusion is that applying social scientific theory to the modelling of social reality seems a promising research avenue.

1 INTRODUCTION

This text complements the NIAS-Lorentz Theme group proposal "emergence of the glass ceiling" (doc to be obtained from author). It constitutes 'thinking aloud' in preparation for the project. The first objective is to find out how Kemper's theory [1] could be related to Hofstede's dimensions of culture [2] for modelling human social behaviour. Kemper and Hofstede provide theories that address the question why people do things. For agent-based modelling, this is not enough: theories are also needed that address the question how. We do this in the light of a sample activity. The aim of this study is to shed light on the emergence of social configurations and institutions at society level, from activities that occur in everyday social life. For the NIAS-Lorentz study we chose one instance of this: the emergence of the glass ceiling for women in organized life. The studies of Hemelrijk et al. in macaque societies show that the violence of dominance interactions can explain both spatial structure and the relation of dominance position and gender [3]. We assume that similar processes could be at work among humans; but the complexity of human society is baffling. Children's play has a lot of the social interaction of life, without the institutional complexity. It seems a fitting research laboratory. We therefore also discuss children's play in agent-based computer models.

This is a work in progress. The theoretical coverage in this document is still far from complete, and so is the integration. Frank Dignum, Rui Prada, Ana Paiva and myself will spend the fall semester of 2013 working on these issues, and invite your thoughts.

The paper is structured as follows. First the generic social scientific theories that will serve as the basis for agent behaviour are introduced. Then I review applied ethnographies of children's behaviour, which is the case that forms the context for agent behaviour. I stress the possible emergence of pattern from child behaviour, and its role in leading to a glass ceiling for women. In the next section I discuss how to use all of these elements in building agent-based simulations, and I present an example in progress.

2 THEORY USED

2.1. Kemper's status-power theory

Theodore D. Kemper (1926) spent his career exploring human social behaviour from a sociologist's perspective. In his 2011 book [1] he sums up the resulting theory. Kemper sets out to explain why we do what we do, more than the details of all the things we could possibly be doing. His theory states that our social behaviour revolves around the concepts of status and power. It could be summarized as "Make status, not power".

Status as Kemper uses it is not just a pecking order variable, though it includes that element. It is something that we continually both claim from one another and confer upon one another through our actions. An example may be the fastest way to explain. If, at the office, I greet X upon entering their room unannounced, I confer status on X; how much will be determined by the modalities of the greeting. My choice of greeting will depend on things such as our hierarchical and personal relationship, what preceded between us, my personality, the nature and urgency of the issue at hand, and whether others are present. At the same time, by entering unannounced I make the status claim of being somebody entitled to enter X's room. Formally, status is the voluntary compliance with the wishes of another. It is a concept akin to Maslow's [4] affiliation (the wish to confer status, or the status that others confer upon one), as well as to his dominance (status as a pecking order variable).

Power comes into play when we want someone to do things and they do not voluntarily comply: we can then coerce them in some way, by pleading, lying or violence. The difference between power and status is a tricky thing; many actions have a power and a status component. For instance in our example, if X does not want to confer status upon me by hearing me, (s)he could look up, say "Excuse me, but I'm, very busy, could you come back later?" and then resume working; this might be a status move, indicating that I have not enough status to enter, or (s)he has to keep working for our common boss. I'd probably also interpret X's action of resuming work as a power move – I would have wanted X to continue looking at me to hear my reason for entering, and I expect X to know this. A blatant power move would be if X stood up, beat me around the head and shoved me out of the room. Obviously, using power too rashly would be unwise for X. If (s)he beat me, I'd probably go around and tell everyone, and the result would be that X's status in the

[1] INF group, Social Sciences,, Wageningen Univ., Hollandseweg 1, 6706KN, Nl. Email: gertjan.hofstede@wur.nl.

wider reference group to which we both belong would be lowered. The only justification for such behaviour on X's part would be that I had previously given X a horrible status affront – for instance, by declaring him/her to be a fraudulent researcher.

Reference group is another important notion in Kemper's theory. Sociologically speaking, he says, we have in our mind a committee of reference groups deciding about our actions. Sometimes this can be quite complex; e.g. when the greeting rules from the football club, where I play in a team with X, differ from those of the office, which ones to use? I might confer more status upon X by saying 'Hi, Zizi' – but other office members might object, and withdraw status from me, if I did that. And those others do not even have to be physically present; all that is required is that they be in our mind. Hence emotions such as pride, guilt or shame that we can feel when quite alone, or the expression 'God forbid'. A sports team, an set of colleagues, a religious community, or a society, these are all reference groups *sensu* Kemper.

Kemper's theory posits that people attempt to maximize their status while protecting themselves from the power of others. This may sound like economic rationality re-invented. But the trick is that status is earned by a proper dose of status conferral upon others, refraining from over-claiming status with them, and using power in ways backed by authority granted by the reference groups. So people are dependent on being upstanding members of their reference groups for obtaining the high status they crave.

The mechanism that helps people take care of their status-power interests is called emotions. Note that when modelling emotions, other sources are also important, notably the encyclopaedic work by Frijda [5]. Kemper divides emotions into three main groups: Structural, situational and anticipatory emotions. He finally distinguishes between technical and social activity. Technical activities have practical goals, such as feeding oneself or building a tree house. They are usually carried out in ways that also serve relational goals, e.g. enjoying one another's company (mutual status conferral).

2.2. Hofstede's dimensions of culture

Geert Hofstede (1928), an engineer turned social scientist, was involved in large-scale opinion surveys at IBM international in the late sixties. The results showed surprising regularities, not across gender, tenure or job type lines but across nationalities. Geert decided to pursue the topic. After ten years of study he came up with a theory [6] that became very influential and much-replicated. The most recent version of that theory, incorporating findings from other studies, is described in [2]; there are now six dimensions of culture in the model, each of which represents one of the big issues of social life that the members of a society have to contend with. These issues are about independence, authority, aggression, anxiety, change and freedom. The associated dimensions are bipolar continua, on each of which each society takes a position. These societal traits are not to be confused (but, alas, often are) with personality traits such as those found by McCrae et al, although there are national-level correlations [7].

2.3 Status, power and culture

Is there a relationship between Kemper's model and Hofstede's dimensions of culture? If social life revolves around status claims and conferrals, and power exertion and avoidance, then this should be reflected in dimensions of culture. We would expect different societies to have different propensities to use power, for instance; power sanctioned by a society being known as authority. Empirical research into this question would be very difficult, since Kemper's constructs apply to real-time interactions between people, whereas Hofstede's dimensions of culture are derived from country-wide central tendencies. We can speculate, however, that the dimensions point to systematic differences in how the people in a culture tend to act – thus both enacting and perpetuating their culture, and sometimes modifying it.

Some support for this position can be found in the fact that at country level, Hofstede's dimensions strongly correlate with the big five personality factors [7, 8] – so correlations between constructs at society level and averages of constructs at individual level can happen, and be meaningful. For instance, the personality factor of neuroticism correlates positively with masculinity and uncertainty avoidance, and I have interpreted this as 'fear' in what follows.

In what follows I present an account of each dimension of culture in Kemperian terms, followed by an example of two countries that differ particularly much on that dimension.

Individualism is about who are the reference groups in the mind of an agent, taken into account when the agent considers if an action is status-conferring. In an individualistic society, these might be several, differing in their reach of control over the agent's mind. They might include heroes, friends, or one's nuclear family members, deities and fiction characters as well. Different reference groups form what Kemper calls the 'reference group committee' and they pull the actor in different directions. In a collectivistic setting, there is likely to be one inclusive reference group, the extended family, clan, or people, that overwhelms the others. Also, a lot more behaviour is scripted from a societal role point of view in a collectivistic setting. Another aspect of individualism is that technical activities *sensu* Kemper are more likely to be the basis for creating a reference group ('task force'), whereas in collectivistic societies, existing reference groups will be the likely group to execute technical activities. A stark contrast on the dimension of individualism is formed by the United States and Indonesia.

Power distance is about voluntary status-accord to others, and granting of authority, based on ascribed characteristics, not on actions. The net effect is that default status-accord in an interaction will be asymmetric: participants will seek to find out their respective status, and if they deem themselves inferior in ascribed status, they will give way. Some status markers are age and gender. Obvious power can also serve as a status marker, and anyone who is obviously powerful may acquire undisputed status. Small power distance stands for symmetric status-accord. Institutions in such a society will prevent accumulation of wealth and power: progressive taxes, democratic elections. For a contrast on this dimension, compare Israel with Russia.

Masculinity is about voluntary status-accord to others based on their performance in competitive settings – in other words, based on their power. It is also about fear that powerful others will use that power against one's own power. The net effect is that people in interaction tend to seek status either by winning competitive sequences (not just fighting or sports, but also having a bigger car, being more elegant, having higher marks, cracking one-upmanship jokes, having more publications…), or by aligning themselves with powerful 'winners' (presidential candidates, deities, sports heroes). The converse, femininity, stands for voluntary status-accord to those who refrain from using or showing power, and a reluctance to use power, or to accept authority when it is enacted in powerful ways. In masculine societies, games will be about winning, while in feminine ones, they will be about participating. For a contrast on this dimension, compare The United Kingdom with Sweden.

Uncertainty avoidance is about fear of the power of others, though not specific others (as might happen in a masculine culture), but generalized anxiety in the face of anyone or anything unknown. It is also about potential status loss for acting in strange ways, since such acting might release these anxieties. Conversely it is about status accord to anything that brings safety, e.g. a known person, boss, or leader; a specialism, expertise. Uncertainty tolerance stands for confidence in one's own power and, as a result, willingness to face the unknown and to trust institutions and generic good sense. Greece and Singapore provide a striking contrast on this dimension. **Long-term orientation** is about renouncing to immediate status claims or conferrals. This happens because each claim or conferral is done with a view to its effect on the potential for status claims at a later time. Short-term orientation, on the other hand, is about taking status conferral in the here and now very seriously. One expects to confer and receive status to the full, regardless of what may happen later – because not doing so would be a great status loss. One zooms in to life at the moment, as it were. This makes moral issues very important, as opposed to pragmatic ones. A strong contrast on this dimension exists between Japan and Iran.

Indulgence is about allowing free-form opportunities for status conferral to oneself or others, including what Kemper calls 'the organism'. The idea is that role prescriptions and rules can be relaxed or forgotten, which leaves room for all kinds of play and indulgence: in play, food, sex, or violence. The opposite, restraint, holds when constraints are taken very seriously, and infringements lead to loss of status. Hence, people are likely to use non-organismic ways of claiming and conferring status, 'sublimating' the organismic ways. Countries at the extremes of this dimension are Mexico and Pakistan.

If these suggestions hold water, it should be possible to use them in agent-based models of social behaviour that use both Kemper and Hofstede. Where Hofstede jr. [9] asserts that everybody plays the moral circle game whatever they do – with localized differences in the unwritten rules -, Kemper adds that this is a status-power game, and he speaks of the 'reference group committee' in people's mind that guides their decisions. Reference group and moral circle boils down to the same thing.

Theodore Kemper informally suggested the following short formulations, that would be perfect for use in models, but might be cutting some corners:

"As I read the six culture dimensions, they are amenable to construal in status-power terms in the following way:

i) Individualism-collectivism is a society's specification for the unit that has the right to claim and receive status.

ii) Large vs small power distance is the willingness to accept status and/or power domination.

iii) Masculinity vs femininity is a preference for either power-oriented or status-oriented social relations.

iv) Uncertainty tolerance is the rigidity with which status-power rules are mandated to be followed.

v) Long vs short term orientation is a matter of how change in status-power rules is accepted.

vii) Indulgence vs restraint is the degree of control over organismic satisfaction (a matter of status claiming)".

3 SEX DIFFERENCES IN CHILDREN'S PLAY

3.1 Ethnographic studies

Our case is children's play. Kemper does not, in his 2011 book, take up the differences between the sexes. There is other research about this though, that could be interpreted in the light of status-power theory to create different boy and girl agents. Lever [10] spent a year studying 181 fifth-grade primary school children in Connecticut (age 8-12) at three middle-class schools. She used four methods: observation on schoolyards, interviews, questionnaires, and diaries. Excluding e.g. TV watching, she found six differences:

- Girls played more with dolls or board games, mimicking primary human relationships, while boys played sports or "war" outdoors.
- Boys played in large groups more often. This is related to the first point. Even girls outdoors played in smaller groups: tag, hopscotch or jump-rope require fewer participants than team sports.
- Boys played more in age-heterogeneous groups, admitting younger boys when the game required more participants
- Girls more often played in "male" games than vice versa. Girls could be used by the boys for a sports team if no boy was available. They would then seriously try to play. When boys joined in girls' games, it was as "buffoons" or to tease, and they were not censured.
- Boys played competitive games more often than girls. If one distinguishes between play and game, in which only the latter have a formal aim and winners, 65% of the boys' activities consisted of games versus 35% of the girls'.
- Boys' activities lasted longer. 72% of the boys' activities lasted longer than an hour, against 43% of the girls'.

Lever interprets her data as follows. First, the ceiling of skill is higher for boys, so that they keep being challenged by their games. Kemper might add that the challenge is also sustained because the boys' games are more politically complex than the girls' games and include power moves. Second, boys were found to resolve their disputes more effectively. They quarrelled a lot but never let it end their games – actually they also enjoyed the squabbles, especially those who were not particularly proficient in the game itself. Kemper would say that status could be gained through these conflict resolution sessions. By contrast, girls played games that avoided ambiguous, conflict-prone situations. And if such situations occurred, as in girls' soccer, the girls tended to argue about fairness and leave. They also had problems deciding on choosing sides, deciding who was captain or even which game to play. Kemper might here say that apparently, creative moves by one girl were interpreted by the others as power moves in the relationship – and not tolerated.

Girls often played in pairs of "best friends" that reached great emotional intimacy, and could be interrupted by a third party, leading to a kind of serial monogamy of best friends. Where girls learned intimate relational skills, boys learned more instrumental relational skills towards 'generalized others'. While these data are obviously situated in place and history, similar differences seem to obtain in the Netherlands in 2012, according to my unguided observations – but this would need checking. Anyway, to cite Lever [10, p. 458]: "…the world of play and game activity may be a major force in the development and perpetuation of differential abilities between the sexes". This is a good justification for studying children's play.

Barrie Thorne [11] also did extensive field work at primary schools in the USA. She found many of the same phenomena as Lever. She adds thoughtful interpretation about how same-age grouping and public visibility enhance gender salience. Games in which there was explicit team formation had more gender separation than games in which individuals could join. She concludes that the *how* of gender separation may be a good place to start looking for the *why*. Thorne also remarks that dominance is important: dominant kids tend to be male, but on occasion dominant girl troupes (11-12 years old, when girls can be bigger and stronger than boys) would roam the playground.

Of course one expects to find a different picture when studying younger children. Martínez-Lozano et al [12] studied 5-6 year-old children at play in schoolyards. They focused on conflict issues, strategies and outcomes. They collected data in Andalusia, Spain and Utrecht, the Netherlands, in both urban and rural contexts. Sex differences they found across these sites, hence regardless of national culture, were:

- Issue: Girls had more conflicts about possession of objects or space, boys about control of play behaviour.
- Strategy: girls negotiated more while boys ordered, argued, suggested, explained, or accused.
- Outcome: boys submitted more frequently, while girls reached compromises.

It is tempting to surmise that these differences are caused by the girls being more focused on the relationship, whereas for the boys, the joint play activity was more important. If this is true, it is in line with Lever's findings in the USA thirty years earlier.

Conflict resolution in play could be a good topic for agent-based modelling, since from a relational point of view, conflicts endanger the aim of achieving status, as well as potentially involving power, so agents will have a drive to avoid or resolve them.

Developmental psychologists Steenbeek and Van Geert [13] developed dynamic and agent-based models of dyadic child play, and validated their models by performing an experimental study with 48 6-7 year-olds in the Netherlands, using sociometric status ('popular', 'normal', or 'rejected', with girls being over-represented in the latter category) as an independent variable; this can throw some light on the issue. Their model, though not explicitly using status-power theory, would easily be interpretable in its terms. It posits four theoretical principles:

1. behaviour is intentional and goal-directed (this also fits the BDI framework for agents; see below)
2. goals represent concerns in the sense of Frijda [5]; these are similar to status and power concerns in Kemper
3. social interaction is a goal in itself (allowing mutual status conferral, Kemper might say)
4. behaviour is affected by non-intentional copying and mimicking (again: conferring status), preferentially of children with high status or power.

The authors found some unexpected things in their empirical study. The setup was that dyads were assembled for the occasion and left to play at a table with a video recorder for fifteen minutes. Later, the sequences were coded according a theory-based model based on the principles above. 'Rejected' children turned out to be more 'other-directed' and 'positive', while 'popular' children showed 'negative expression' more often. The status-power explanation might be that, since the popular children were not in danger of walking off to find other popular children, the 'rejected' ones were free to try and make themselves loved – acquire status – while the popular ones could punish (refuse to accord status).

Another feature of the study that merits consideration for the present proposal is that not only single playing sessions were modelled, but there was also a model about the long-term development of patterns that could result from repeated interactions. This methodology allows to get a grip on emergent patterns of behaviour.

A recent study by Tessa Lansu [14] on Dutch 10-12 year olds showed that children who professed liking dominant other children really showed repulsion to those others in subconscious responses. In particular, aggressive girls really disliked powerful others, contrary to boys. Popular girls were sensitive of others' needs, again contrary to boys. This puts into question the relationship between 'liking' or status-accord, power and fear in the context of children's groups. Since a similar pattern of avoidance of power by girls, in this case rough-and-tumble play, was found in pre-schoolers in the USA [15] it seems to be a robust finding across age and culture, that might explain some of the patterns found in ethnographic studies.

To conclude: the topic of child play, with age, sex and relative status-power as conspicuous variables and across different time scales, seems a promising topic for the study, both from a theoretical perspective and from a practical one.

3.2 Emergent structures and children's play

Children's play is widely interpreted as preparing for later life. The idea is that children play games in which they learn to enact roles and keep to certain social configurations. Being a boy or a girl is usually believed to be important for children's play. The studies mentioned above support this idea: there is a clear differentiation between play by boys and girls. Traditionally, the debate about why men and women behave as they do has been caught in the nature versus nurture debate. Recently, self-organization, leading to emergent structures, has been added [16]. Hemelrijk puts it thus ([16], p.224): "It appears that the discovery of cognitively simpler explanations is furthered by the use of self-organization models". Whether and how self-organization leads to emergent results at society level in adult life will be a major question in this research. There could be three perspectives on this question: nature, emergence, or nurture.

- Nature: Boys and girls are biologically different, and therefore go on to play different roles is whatever society.
- Emergence: There are regularities about status and dominance interactions between boys and girls, that scale up to adult life, leading to patterns and institutions in society.
- Nurture: It is the pressure of adults and peers that lead boys and girls along different developmental paths.

The nature hypothesis, while undeniably containing truth, does not do well in explaining cultural differences and societal changes. The other two are needed for filling out these details. This study can help uncover whether the emergence perspective can explain part of the picture.

What are the things that might emerge? Hemelrijk's [16, 17] studies can be an inspiration here. Based on a single variable, 'despoticism', she found emerging gender role division and spatial structures in macaque populations. Hemelrijk started with a population of 10-12 individuals who were all of equal status, except for sex differences (males being larger) that could vary across macaque species. Then these individuals engaged in interactions as in fig 1 – and in nothing else. Despoticism modelled the violence of dominance interactions – in other words the amount of power used – and the effect on scalar status of winning or losing such an interaction. This makes it a concept akin to the cultural dimension of masculinity / femininity. Unexpected outcomes lead to more status change, both for their winners and their losers, than do expected outcomes. The aggression level of an attack also changes the status effect. In more despotic macaques, e.g. rhesus monkeys, male losers of fights lost so much status and felt so bad that they moved to the outside of the group, in order to get away from their vanquishers. This led to a typical spatial structure known from baboon rocks in zoos where the dominant males occupy the centre, and the subordinate animals the periphery. Females, being smaller than males, would be numerous on the outskirts of the group. Males and females thus lived in different social spheres. Recent male losers who had fled from the centre were subsequently vulnerable to further loss – also against the females who could be found there, and some of whom might feel really dominant after just having beaten other females. Actually, the smaller the power of females compared to males in the model, the steeper the hierarchy among females would become, and the more despotic females would on average become. At the 'egalitarian' (in Hemelrijk's terms), feminine (in Hofstede terms) end of the scale, stump-tailed macaques resolve their dominance issues by staring one another down, without resorting to power use such as biting. As a result, losers did not lose much of their scalar status and did not move away, individuals of both sexes would mix in a rather homogeneous way, individuals would live closer together, attack one another (that is, stare down) more often, yet no females would ever rise above any males in the status hierarchy. In the model Groofiworld, in which individuals would groom those whose power they feared, egalitarian groups would show more grooming – also because they were freely mixed and would often encounter others they feared. An interesting side thought is that these egalitarian, status-conferring groups might also constitute a natural laboratory for theory of mind since everyone is continually at close quarters with people who might have to be groomed.

If similar, but more complex dynamics obtain in humans, then one would expect male and female social circles to be more separate in more masculine societies, with males typically in the powerful roles. Yet one would also expect to have relatively more, rather than fewer, powerful women in these masculine societies – ceteris paribus. DomWorld produced relatively more female dominance in those groups (of size 10-12, in the model) in which the hierarchy was steep and/or females were scarce, and empirical results among primates confirmed this – this might be another phenomenon that occurs in humans, especially in masculine societies. Men fighting one another down around rare women at the top – it sounds like something familiar.

Supposing that the expected differences at society level between Great Britain and the Netherlands can be found, how will this inform our simulations? It makes sense to expect that a lot of adult social structure is formed, or at least some of its primitives are formed, during childhood. For instance one would expect more overt antagonism and power use in GB than in NL, and less boy-girl interaction, with boys occupying central spaces on the playground. At the same time one would expect certain girls to be dominant over certain boys more often in GB than in NL. And probably, these things would vary considerably in relation to development, notably puberty. We can safely expect that sexual attractiveness will increase boys' conferral of the status to girls who 'meet standards' as Kemper puts it, and this will affect older children more than younger ones. The age at which, and extent to which sex issues change the game probably also change with social class, depending on future prospects; if one expects a long life, then sex becomes less important [18].

In order to discover these emergent effects, it is preferable to study groups of children in free play rather than just dyads. Emergent effects could be: in masculine culture, children might use more space, have steeper hierarchies, have less spatial overlap between boys and girls with boys at centre space, have fewer power-free status exchanges. A modification could occur if competitive games come into the picture: they could be a legitimized form of power use while conferring status on all participants, and hence a very successful activity in masculine cultures.

4 SIMULATING

In order to be able to simulate, any ambiguities in theories used, or conceptual holes between them, must be resolved. Here are some preliminary ideas that might be usable.

4.1 Simulating children's play

We shall try to build models that are as simple as possible, while still allowing to grasp cross-cultural variation in social behaviour. Kemper's parsimonious model makes it possible to build models of social behaviour in which agents are driven by status-power considerations. Our ambition now is to look at social behaviour in the absence of an economic decision. We choose the subject of children's play. This topic is chosen for its obviously social and relational nature, without ulterior economic motive [19], for its assumable role in the emergence of social, culture-bound patterns, and also because there are empirical data available, or obtainable, across cultures.

In Hemelrijk's models [16, p. 225], status-power dynamics also obtain. Agents are driven to claim status with or without use of power, if they meet an individual whom they think is weaker; else, they confer status by grooming if they fear the other's power. Would we find similar dynamics in children? Complementarity Theory by U.S. anthropologist Alan Page Fiske [20] posits that there is a co-evolution of innate psychological capacities in children with culture-specific co-ordination devices; this is exactly what we would like to grow in our models, and it reminds the 'how' question of gender salience put by Barrie Thorne. Fiske puts it as follows (p. 76):

”Putting proclivities together with congruent paradigms, children learn to construct culture-specific coordination devices that enable them to interact in locally meaningful ways. The evolved proclivities and cultural paradigms are complementary: Both are necessary but neither is sufficient to permit complex social coordination.”

Kemper's theory can provide the essence of our 'evolved proclivities', while Hofstede's dimensions can provide the core of cross-cultural variations. We should this be able to use them to create an agent-based Complementarity Theory laboratory. Kemper's theory can be used in agents' state variables about the status they believe they have with others, and the status they believe others are claiming, and the status they are willing to confer on others. They need state variables about the power they believe they have, and their fear of the power of others. Hofstede's theory can be used as salience mechanisms and filters various element of agents' social functioning.

For these 'various elements' in agent-based models, theories about the mechanism of interaction are needed. Agents need to perceive one another's actions, interpret them, and respond. They need a memory too. Current architectures for social interaction among agents are often based on two frameworks [21]: the Beliefs-Desires-Intentions (BDI) model for agent minds [22], and the Ortony-Clore-Collins (OCC) theory for emotional appraisal by agents [23]. Both models can be used with the theories presented here. OCC includes 'praiseworthiness' and 'blameworthiness' and these are directly translatable to status conferral and withdrawal à la Kemper.

A micro-theory about the simulation's topic, i.e. about what actions play involves, will be needed too, and we shall consider that.

Several kinds of simulations with different level of detail can be considered. In detail simulations, play sessions with set rules can be modelled – possibly with embodied agents [24], but not necessarily - and the emergent results collected across runs. In zoomed-out simulations, the emergence of co-ordinating devices such as norms and rules [25] can be studied. Simulations can be made that allow rules to change across time, or across groups. It is quite conceivable that in children's play there are punctuated equilibria, in which new groups are much more likely to come up with new patterns than existing ones. This would e.g. appear from the work of Ballato [26] on risk behaviours in adolescents.

4.2 Simulating with Kemper's model

We saw that according to Kemper people, hence also agents in models that use his theory, are dependent on their reference groups for obtaining the high status they crave. This leads to complex emergent status results for group interaction sequences. A game might be enjoyable, with everyone gaining status, or it might lead to a fight resulting in bad feeling. In status terms, this would mean that everyone in the reference group loses status, which could in the long run lead to dissolution of that group.

In play, usual constraints and obligations are suspended to allow maximum mutual status conferral and absence of power use. The only power that can and should be freely used is creative power.

My own surmise is that children are only free to play when they have a comfortable level of status in the group, and are free from fear of the power of others -in other words, if they feel safe. This is for instance nicely shown in the book Momo by Michael Ende [27]. Momo is a little foundling girl who lives in an abandoned amphitheatre, where both children and grown-ups like to come visit her. The other children can play better when Momo is part of the group, and why? Because Momo confers a lot of status and never uses power. She can listen so well that all who meet her end up feeling confident and happy with themselves.

In Kemper's terms, if we consider children in dramatic play, four types of moves are possible:

- Status claim, e.g. "I want to be the queen!"
- Status conferral, e.g. "You be the king."
- Assertion of power, e.g. "If I cannot be the queen, I will not play."
- Accepting other's power: "All right, come back and play. You can be the queen."

Note that each action of any player could belong to more than one of these types at the same time, and that what B perceives could differ from what A intends. For instance, after A has introduced the dragon, if B interprets it as a power move, B could say "And here comes the dragon-drowning cloud, raining down upon the dragon and quenching its fire!" This confers status upon the A for taking up the dragon idea, but also takes status away from A's dragon – which might be seen by A as another power move in the real world, or just as a play move. It also supposedly confers status upon B for coming up with such a clever idea.

4.3 Simulating with Hofstede's model

The Hofstede model comes into the story once a generic model functions. The agents can then be assigned to cultures, and the culture dimensions can be used in decision functions to modify the state variables and behavioural tendencies. Hofstede's model has already shown to be amenable to modelling cross-cultural behaviour in such a manner in agent-based models [28, 29]. These models, however, take a very simple situation as their focus: a one-on-one negotiation in the former paper, and a single-person decision in the latter. Both involve economic decisions about buying. Neither of them zooms in to the relational details of the topic.

In a model of children's play conflict resolution similar to Martínez-Lozano et al.'s, one could model culture by having cultural meta-norms [30] modify the decision functions about conflict perception, strategy, and outcome. Cultural meta-norms specify behavioural tendencies given a certain relational situation. For instance, in the Martínez-Lozano study, Dutch children ended their conflicts more often by walking away, while Spanish children preferred to submit. The authors explain this through the differences in Individualism in the national cultures (Nl: 80, Es: 51). In agents, the individualism score can be used as a modifier for the cultural meta-norm about how to renounce to having it one's way in a conflict, i.e. in this case, the decision to walk away or submit. What the authors do not say is that the dimension of power distance could also contribute to explaining the outcome. With PDI=57, Spain would see more submission than the Netherlands (PDI = 38).

5 META-MODEL FOR SOCIAL BEHAVIOUR

5.1 Rich model

I shall now think ahead about possible model elements. This is really work to do during the project, and it will be dependent upon further decision about the scope of the subject matter. For instance, free play - in which there could be fission / fusion of groups across sessions, very relevant to human social life - is conceptually more complicated than game play – in which there are explicit group boundaries and rules. Larger groups are more complicated than dyads. We also depend on the empirical data that we can get our hands on.

Agent classes

A possible meta-model for the structure of agents has to take into account reference groups as well as individual agents. Moreover, people tend to use not just other people, but also other entities as sources of status conferral or claims, or of power: symbolic entities, groups, heroes, fictitious characters, deities.

A first try could be to use a meta-model akin to AGR [31]. In our model, an agent would be a status-power actor – and this would really be a supertype of which various kinds of persons, entities and groups could be subtypes. A Role would indicate the relationship between two status-power actors, and have scalar status and authority (that is, power sanctioned by the group) attached to it; it could be group membership (between an individual and a group), or friendship (between two individuals), or a formal, authority-carrying role (mother, president, secretary). Roles could have norms attached to them, about e.g. authority to use certain power moves that comes with a role.

Agent drives and emotions

Agents will be driven by status-power considerations according to Kemper, that is they strive to achieve status (possibly using power), confer status on those who meet their standards, and avoid others' power. For brevity's sake I skip this elaboration.

Interaction dynamics

In order for any model to run, agents will need to behave. Other agents will need to perceive, interpret, and reciprocate. Each action will be interpreted by the acting agent, as well as by other agents, in status-power terms, and the interpretations could differ – particularly if the agents have different cultures. The agents' actions could also have a practical side that feds into the status-power dynamics.

Which behaviours to use will depend on empirical factors. Studies on children's plays can be of use here.

An agent-based model of children's play should minimally contain the four types of intentions / interpretations of moves (status claim, status conferral, power assertion, power acceptance / refusal), without necessarily containing any worked instances – that is, there does not need to be a physical or verbal form for the moves. Each child agent should have state variables for its status with the others, and its fear of the power of the others. These variables could be dyadic, even in larger groups. Group-level variables could emerge when there is sufficient alignment between the dyadic variables, so when the group members agree among themselves about status. Children should also have propensities to making each of the types of moves depending on their status and power state variables. There also needs to be a closure criterion for joint activity. If the simulation is about child play, when the total mutual status conferral drops below a certain level, participants will leave the game; otherwise they might stop through external causes, e.g. when they reach an exhaustion threshold.

The ideas presented in this section are by no means the only conceivable way to instantiate Kemper's and Hofstede's theories. The role of empirical evidence will have to grow. There can be many different simulations of various games, settings, age groups. The counts-as operator can map instantiated concepts from those various simulations to the status-power primitives that form the basis of our work.

5.2 Minimal model in Netlogo: Playground

To give a first impression of how Kemper and Hofstede can be used, without instantiating any behaviour beyond the theoretical concepts, I developed a small model in Netlogo 5.0.4. (figure 1). It uses the theory above. No validity testing has happened at the time of writing, so the current version is based on literature only. The model has been kept extremely simple, and will only be made more intricate if the simpler model does not yield any emergent result. For the moment the children have no relational memory and no theory of mind, and minimal emotions. The main logic runs as follows: Agents are boys and girls. They have two Kemperian variables: beauty, and power. Beauty makes one more likely to receive status, while power is used in quarrels. All agents look for one or more friends and then start to perform status exchanges. The only difference between boys and girls is that girls avoid exchanging status with more powerful others, in line with the findings in [14] and [15]. If they feel short-sold on status, they may start a quarrel. Power, beauty and culture parameters modify these exchanges.

A *status exchange* during play involves two children in the same group (shown by colour) conferring status upon one another. This models a micro-interaction within the group process, such as one child playing the ball to another, smiling at it, or racing it, and the other interpreting that action as appropriately nice, more so, or less so. The dynamics of the Kemperian exchanges are modified by Hofstedian culture parameters. *Large Power distance* makes a child gauge the status conferral by relative power, so that it expects larger conferrals from less powerful others. Note that so far we do not use a separate measure of social status here, only physical power; the playground is not a multi-class place. *Masculinity* makes a child less tolerant of a status deficit.

```
ifelse status-conferral + (power-distance * power) >
    mate-status-conferral + (power-distance * your-power) +
        ((100 - masculinity) / 33)    [ get angry ]
```

An angry child may pick a quarrel, which is a Kemperian power exchange. A quarrel is a comparison of both children's power, and thus the stronger sex (determined by a slider) are more likely to win fights than girls. The winner plays on with the group and gains a little power. The loser becomes unhappy and loses a little power. These power changes vary with masculinity – modelling level of fierceness of the fight. If a too large fraction in a group are unhappy, the group dissolves.

So far, an emergent effect of power distance, in combination with children's power, is to make fighting more likely, since exchanges become more asymmetrical. Masculinity also leads to more fighting. The corner testing so far seemed to indicate an emergent gender-related pattern of boys being slightly happier when fights occurred; this is due to their greater average power. More interesting, and reminiscent of Hemelrijk's work, is the effect of girls' avoidance of others' power. This leads to boys starting more fights and being more likely to become dominant (increase their power), even if girls are stronger on average.

Of course, at the time of writing, the above is no more than a small first step. The fact that emergent gendered patters occur even if girls and boys do not consciously avoid one another is promising. Kemper seems very useful as a generic model that can be instantiated for various cases – in this case, child play. Hofstede has the same properties of being generic and instantiable, but national culture is a population-level concept. Ideally, culture as a system-level property should emerge from agent-based models rather than being a set of input variables; but then we are speaking of evolutionary models. Creating these is a tall order. In a model that does not span across generations, national culture can be assumed to be constant [2] and thus be used as a set of parameters. If the basic Kemperian process dynamics are well modelled, thoughtful incorporation of ethnographic findings like [12] from several cultures could allow calibrated use of culture sliders. Face validation of the model by experts from various cultures is also possible.

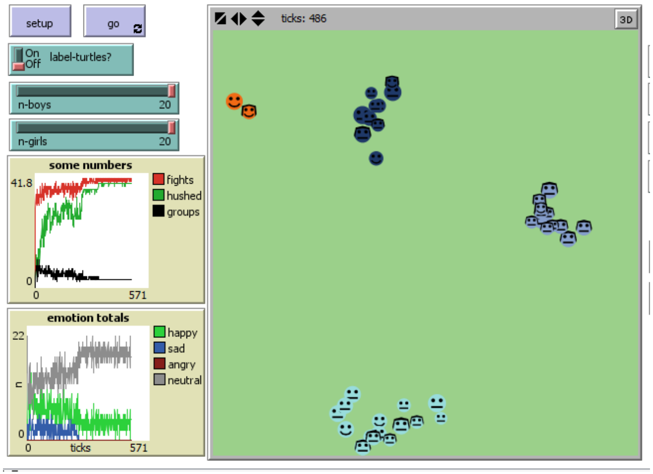


Figure 1: partial window of a sample run with culture parameters set to 50. With these settings, some groups form in which fights are hushed. Children fight initially, losers are sad and run off. Gradually, stable groups form and fights are hushed, leading to neutral, subdued faces. The run shows one remaining unstable pair happily playing (top left).

6 ACKNOWLEDGEMENTS

The ideas put forward here have greatly benefited from discussions with my colleagues in the projects eCute and SEMIRA, by Theodore Kemper's comments, and by anonymous reviewer comments.

7 REFERENCES

[1] Kemper, T.D.: 'Status, Power and Ritual Interaction; A Relational Reading of Durkheim, Goffman and Collins' (Ashgate, 2011. 2011)

[2] Hofstede, G., Hofstede, G.J., and Minkov, M.: 'Cultures and Organizations, Software of the Mind' (McGraw Hill, 2010, 3rd edn. 2010)

[3] Hemelrijk, C.K.: 'The use of artificial-life models for the study of social organization', in Thierry, B.S., M; Kaumanns, W (Ed.): 'Macaque Societies: A Model for the Study of Social Organization' (Cambridge University Press, 2004), pp. 295-313

[4] Maslow, A.H.: 'Motivation and Personality' (Harper & Row, 1970, 2nd edn. 1970)

[5] Frijda, N.H.: 'The Emotions' (Cambridge University Press, 1986. 1986)

[6] Hofstede, G.: 'Culture's Consequences, International Differences in Work-Related Values' (Sage, 1980. 1980)

[7] Hofstede, G., and McCrae, R.R.: 'Personality and Culture Revisited: Linking Traits and Dimensions of Culture', Cross-cultural Research, 2004, 38, (1), pp. 52-80

[8] McCrae, R.R., and Costa, P.T.j.: 'Personality in Adulthood: A Five-Factor Theory Perspective' (Guildford, 2003, 2nd edn. 2003)

[9] Hofstede, G.J.: 'The Moral Circle in Intercultural Competence: Trust Across Cultures', in Deardorff, D.K. (Ed.): 'The SAGE Handbook of Intercultural Competence' (Sage, 2009), pp. 85-99

[10] Lever, J.: 'Sex Differences in the Games Children Play', Social Problems, 1976, 23, (4), pp. 478-487

[11] Thorne, B.: 'Gender play; girls and boys in school' (Rutgers University Press), 1993

[12] Martínez-Lozano, V.S.-M., José A.; Goudena, Paul P.: 'A Cross-Cultural Study of Observed Conflict Between Young Children', Journal of Cross-Cultural Psychology, 2011, 42, (6), pp. 895-907

[13] Steenbeek. H. and Van Geert, P. ' The Empirical Validation Of A Dynamic Systems Model Of Interaction; Do Children Of Different Sociometric Statuses Differ In Their Dyadic Play Interactions?' Developmental Science, 2008, 11(2), 253-281.

[14] Lansu, T 'Implicit processes in peer relations: Effects of popularity and aggression' PhD thesis, Nijmegen University, 2012.

[15] DiPietro, J.A.: 'Rough and tumble play: A function of gender.' Developmental Psychology, Vol 17(1), 50-58 (1981)

[16] Hemelrijk, C.K.: 'Simple Reactions to Nearby Neighbors and Complex Social Behavior in Primates', in Menzel, R.F., J. (Ed.): 'Animal Thinking: Comparative Issues in Comparative Cognition' (MIT Press, 2011), pp. 223-238

[17] C.K. Hemelrijk ' The use of artificial-life models for the study of social organization.' In: B. Thierry, M. Singh & W. Kaumanns (eds): Macaque Societies. A Model for the Study of Social Organization. Cambridge University Press, 2004, 295-313.

[18] Wilson, D.S.: 'Evolution for Everyone' (Bantam Dell, 2007. 2007)

[19] Huizinga, J.: 'Homo ludens' (Tjeenk Willink, 1938, 1952, 4th edn. 1938)

[20] Fiske, A.P.: 'Complementarity Theory: Why Human Social Capacities Evolved to Require Cultural Complements', Personality and Social Psychology Review, 2000, 4, (1), pp. 76-94

[21] Hofstede, G.J., Mascarenhas, S., and Paiva, A.: 'Modelling rituals for Homo biologicus'. ESSA 2011 (7th Conference of the European Social Simulation Association), Montpellier, 19-23 September 2012 2011

[22] Georgeff, M.P., and Ingrand, F.F.: 'Decision-making in an embedded reasoning system', in Editor (Ed.)^(Eds.): 'Book Decision-making in an embedded reasoning system' (1989, edn.), pp. 972-978

[23] Ortony, A., Clore, G., and Collins, A.: 'The Cognitive Structure of Emotions' (Cambridge University Press, 1998. 1998)

[24] Dias, J., and Paiva, A.: 'Feeling and reasoning: A computational model for emotional characters', in Carlos Bento, Amílcar Cardoso, and Dias, G. (Eds.): 'Progress in Artificial Intelligence' (Springer, 2005), pp. 127-140

[25] Dignum, F., Morley, D., Sonenberg, E.A., and Cavedon, L.: 'Towards socially sophisticated BDI agents', in Editor (Ed.)^(Eds.): 'Book Towards socially sophisticated BDI agents' (2000, edn.), pp. 111-118

[26] Ballato, L.: 'If I like you, I wanna be like you!' (Ridderprint B.V., 2012. 2012)

[27] Ende, M.: 'Momo' (Thienemann Verlag, 1973. 1973)

[28] Hofstede, G.J., Jonker, C.M., and Verwaart, D.: 'Cultural Differentiation of Negotiating Agents', Group Decision and Negotiation, 2010, 21, (1), pp. 79-98

[29] Roozmand, O., Ghasem Aghae, N., Hofstede, G.J., Nematbakhsh, M., and Baraani, A.: 'Agent-based modeling of consumer decision making process based on culture and personality', Knowledge-Based systems, 2011, 24, (7), pp. 1075-1095

[30] McBreen, J., Di Tosto, G., Dignum, F., and Hofstede, G.J.: 'Linking Norms and Culture', in Editor (Ed.)^(Eds.): 'Book Linking Norms and Culture' (IEEE, 2011, edn.), pp. 9-14

[31] Ferber, J., Stratulat, T., and Tranier, J.: 'Towards an Integral Approach of Organizations in Multi-Agent Systems', in Dignum, V. (Ed.): 'Handbook of Research on Multi-Agent Systems' (IGI Global, 2009), pp. 51-75

An Idea for Modelling Group Dynamics in Autonomous Synthetic Characters

Naziya Hussaini, Ruth Aylett[1]

Abstract. This paper discusses an approach to more believable behaviour by groups of synthetic agents. It considers the theories of William Schutz based on interpersonal relations and provides an initial discussion how it might be implemented within the FAtiMA-PSI architecture.

1 INTRODUCTION

Much research in the field of synthetic characters has been carried out in recent years, which results in developing autonomous believable agents. For example, various conversational agents like GRETA [11] who communicate with users and assist them in their queries. Such agents have influenced various domains of interactive narrative, educational, and persuasive systems. 'FearNot' and 'ORIENT' are such examples. FearNot (Fun with Empathic Agents Reaching Novel Outcomes in Teaching) [6], a virtual interactive drama, in which emergent narrative approach, was applied in synthetic characters to teach children about bullying by letting them interact with an emotional character. ORIENT (Overcoming Refugee Integration with Empathic Novel Technology) [9, 13], an interactive role-playing game, based on the intercultural communication was design to teach children about cultural differences.

Persuasive systems are another development where agents interact with user through conversation and try to convince them by influencing their attitude and behaviour [10].

Autonomous synthetic characters are now able to plan, and execute their own actions and are also able to react to users and other characters affectively. With having such capabilities and autonomy, the next major development would be to make the autonomous characters to act and react in a way that they simulate the group behaviour and are able to have emotional reaction in groups. It will increase their social interaction with each other and thus increase their believability in groups. In a group, it is important that agents should be able to come together with their respective priorities or individualities and manage to have a comfortable relation with each other. The 'coming together' may be for the reason of having lunch together, or an informal play group on weekends or anything which is less formal in nature and based on personal volition. The common goal in such involuntary group would include anything that helps in coming together and spending some time. However, even such common goal of the group should be able to influence their behaviour responses and activities to an optimum level to achieve the goal with less effort and in less time.

The artificial characters that can show emotions and learn from their previous mistakes, if enhanced with group dynamics, can improve their believability at group level. These characters can make a group, which has its separate identity, and features that could be influenced by individual characters but are different from their identities and features.

This paper proposes an idea of modelling group dynamics in autonomous synthetic characters. In the next section, we will start with the discussion on group dynamics. In the subsequent sections we will discuss about the theories related to it and then about the different agent architectures. The paper will try to limit its scope to small informal groups, which will not have many rules and the focus of the study would remain restricted to improving the believability of the artificial synthetic characters.

2 BACKGROUND

Numerous thinkers like, Gustav Le Bon [1], Sigmund Freud [2], Kurt Lewin [3], William Schutz [4] studied the social aspects of the human beings. They tried to analyze different types of groups for different reasons ranging from economic development, decision-making, team building, trainings, management etc. Their purpose was to understand and conceptualise how people interact with each other in groups and the factors that influence their behaviour. These factors could be a lack of belonging to the group or lack of affection among the members of the group [4]. We argue that if these factors could be parameterised, they could be used to reorient the dynamics of the group of synthetic characters. Study of group dynamics can help in tangible ways to optimise processes or improve productivity or just satisfaction of the members of the group.

2.1 Group Dynamics

Group dynamics have played a very influential and critical role in the development of humans. In general, group dynamics refers to the study of formation of groups, the behaviour of individuals, interaction between members within the group (intergroup interactions) or between the groups (intragroup interactions) and the group resolution [4, 12].

[1] School of Mathematics and Computer Science, Heriot-Watt University, Edinburgh, UK.

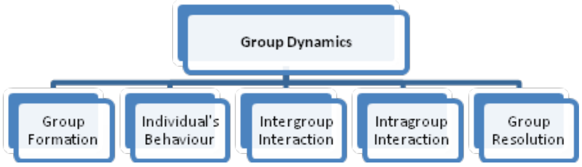


Figure 1. Group Dynamics

There were many theories of group dynamics developed since the days of Gustav Le Bon (The Crowd: A Study of the Popular Mind, 1896) to Richard Hackman (Leading Teams: Setting the Stage for Great Performances (Book), 2002); William Schutz's theory of interpersonal relation is one of them. His theory and the instrument (FIRO-B) based on it were very popular and widely used in groups training.

2.2 Theory of Interpersonal Relations

In 1958, William Schutz introduced a theory, which is based on interpersonal relations known as Fundamental Interpersonal Relations Orientation (i.e., FIRO Theory, rhymes with 'Cairo') that mainly deals with the activities that take place in a group starting from the formation of a group until the objectives of the group are achieved.

According to the theory, people in a group interact on the basis of three interpersonal needs: inclusion, control, and affection, which helps them to develop an interpersonal (or group) behaviour and helps to form a healthy relation with other members in a group.

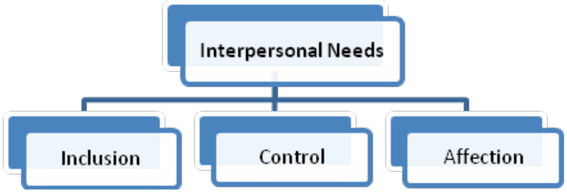


Figure 2. The Interpersonal Needs

2.2.1 The Interpersonal Need for Inclusion

Schutz defined the interpersonal need for inclusion as a behavioural need to establish and maintain satisfactory relation with people with respect to interaction and association. Inclusion is much more than participation. It requires that the different views of the participant be engaged and appreciated. It can be explained, in terms of a computational model, in which every synthetic character with its different, although defined capabilities would try to engage in the group activity through interaction. The other synthetic characters would try to respond to such interactions in order to adjust in the group.

The flip side of inclusion is that it may involve endless interaction within the group and may result in losing sight of the main objective. The means to achieve the objective may themselves become an objective. Thus, the need for control arises.

The need for inclusion mainly deals with the problem of being part of the group (in or out).

2.2.2 The Interpersonal Need for Control

Once we establish inclusion in the group, it needs control and direction. Control, as opposed to the dominance of a member or the sub-group of members, refers to the acts that provide leadership to the group. In the process of inclusion where everyone is trying to propose and contribute to the common task, there are good chances that it may lead to an anarchic or chaotic scenario, which lacks discipline. Leadership and control would be required in such situation to manage the process of interaction and to take effective steps to achieve the group objectives. This may be achieved by defining a member of the group as a leader or monitor who may assess and try to ensure that proper progress is made in a given timeframe. Parameters should be defined in a manner that such leadership be derived from positive factors like being reasonable, trying to focus on the main objective.

The interpersonal need for control satisfies the need for competence. It mainly concerns with the position of members in the group (top or bottom).

2.2.3 The Interpersonal Need for Affection

Finally, after resolving the issues, the need for affection becomes prominent. According to Schutz, "the interpersonal need for affection is defined behaviourally as the need to establish and maintain a satisfactory relation with others with respect to love and affection." [4]

Emotions will start playing their role; each member would try to attain both a closer relationship to seek affection and distance to avoid any unpredictable clash, with the group members. This helps the person to feel that he/she is 'likable'. It deals with the feeling of being close or distant.

2.3 The Postulates of Group Development

According to Schutz, the formation and development of the interpersonal relation (that is, a group) of two or more people always follows the same sequence [4].

2.3.1 The Principle of Group Integration

The interpersonal relations follow a sequence starting from inclusion and followed by control and affection. This cycle may repeat itself until the termination of the group.

2.3.2 The Principle of Group Resolution

The cycle discussed above while integration of the group would reverse itself at the time of termination of the group. The interpersonal behaviour at the termination phase will be more in the area of affection and followed by control and inclusion. The cycle can be presented in the following figure where 'I' stands for Inclusion, 'C' stands for Control and 'A' for Affection. Initially it starts from I and followed by C and A. The pattern keeps repeating itself until just before termination. At the point of termination the pattern will reverse i.e. A, C and followed by I.

I – C – A – I – C – A – I – C – A.........A – C – I

(Where, I = Inclusion, C = Control, and A = Affection)

This theory helps to develop various qualities such as team building, togetherness, decision-making, etc. Since it has a formal structure, therefore this theory can be suitable in defining group parameters in the artificial characters.

2.4 Computational Model of Synthetic Characters

There are different architectures that are based on emotions and are designed to create autonomous believable characters such as FAtiMA, PSI, and FAtiMA-PSI.

2.4.1 FAtiMA

FAtiMA, i.e. FearNot Affective Mind Architecture [5], was based on the cognitive structure of emotions as described in OCC [Ortony et al 88] cognitive theory of emotions. The OCC (Ortony, Clore & Collins) model specifies 22 emotion categories and on this basis emotions were generated in the synthetic characters. FAtiMA architecture was design to create agents with attitude or we can say autonomous characters and is the extended version of the architecture used in VICTEC project [6]. It consists of an autobiographic memory, which helps the characters to remember past events, and on that basis, further action is taken. It also consists of two layers- reactive layer and deliberative layer. Reactive layer is responsible for the character's emotional reactions and reactive behaviour or in short the impulsive action quickly taken by the character without thinking it while the deliberative layer achieves the character's goals by planning appropriate actions.

When an event occurs, the reactive layer checks the emotional reaction rules that were already defined in the memory on the basis of OCC theory of emotions and finds ways, to react in that particular situation. These emotional reaction rules define the values of OCC variables (i.e., desirability and praiseworthiness of events and actions) which further help to generate emotions in a character. After generating emotion in a character, action tendencies provide certain action rules to act in that situation.

The deliberative layer follows the planning mechanism and generates emotions based on intentions. The intention/goal with a high value of success is selected and the planning is done to achieve it. There are two types of goals: active-pursuit goal and interest goals.

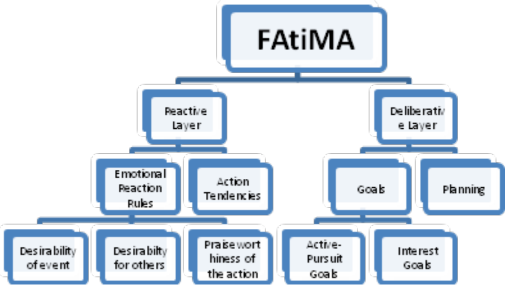


Figure 2. FAtiMA Architecture

2.4.2 PSI

There are other models like PSI theory [14,15,16] of emotions by Dietrich Dorner in which emotions emerge from cognitive processes and emotional parameters rather than predefined in categories as in OCC. It includes different approach from FAtiMA as working of this model depends on the drives of the characters and their thresholds for their particular needs, and as a result the emotions are emergent by different values of the emotional parameters. The PSI model has been used in many researches such as Culture-Personality based Affective model [7] and Affective tour guide system [8].

2.4.3 FAtiMA-PSI

The architecture FAtiMA-PSI, as the name suggest is the combination of both the architecture FAtiMA and PSI. It overcomes the problems of psychological plausibility and control faced by both the architecture, which cannot be solved by either of architectures on its own [9]. It is the modified form of FAtiMA in which the concept of drives (Energy, Integrity, Affiliation, Competence, and Certainty) from PSI was included which reduces the work of reactive layer.

Since the event, actions, and goals are influenced by the drives therefore, there is no need to define everything, agents would learn by themselves, which helps to provide an easy and flexible mechanism. Therefore, we have chosen FAtiMA-PSI architecture for modelling group behaviour in synthetic characters.

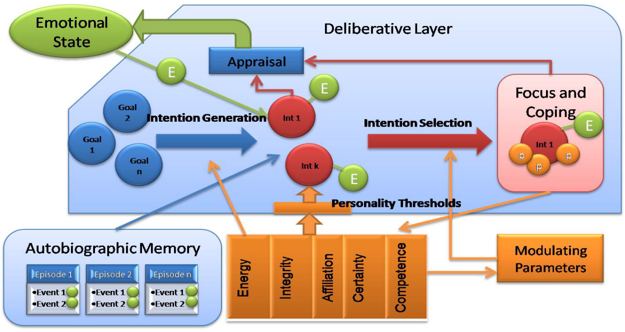


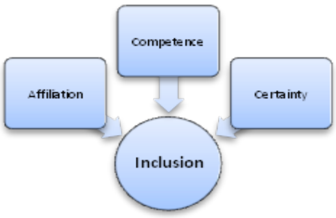
Figure 3. FAtiMA-PSI Architecture (From [9])

3 IMPLEMENTATION

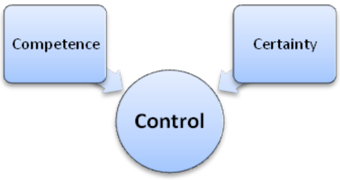
The implementation would include the use of William Schutz's theory in FAtiMA-PSI model. Due to paucity of time and space, the implementation would be strictly to achieve the objectives of this study i.e. to increase the believability factor of the autonomous synthetic characters in groups.

We argue that group dynamics can be modelled by changing the need thresholds of the motivation part of the FAtiMA-PSI model. The group parameters would influence the drives of the architecture. A mapping would be done between the group parameters and the drives. As the value of these drives changes the need for inclusion, control and affection would generate which results in forming group.

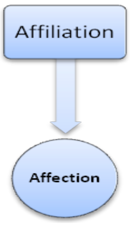
Inclusion dimension can be implemented using affiliation, competence, and certainty. For example, the value of uncertainty avoidance will be high when the group members are not happy with the agent and he has to leave the group, he will not be sure what will happen now and how his task would complete. This causes the affiliation and competence level to go down and thus increases his need for inclusion, which helps him to form a group.



Similarly, Control dimension can be modelled using competence and certainty. If the value of competence were high then the agent would be confident and more certain in performing his task. Thus, help him to have control over the group and would lead the group.



Lastly, the affection can be implemented through affiliation drive. If other members in a group like the agent then the value of its affiliation drive would be high and the agent will feel more affectionate.



A common parameter should be assign to a group whose value ranges between say, 1 to 5. Each character has individual value for this parameter to be a part of the group. Thus helps in implementing the group behaviour in autonomous synthetic characters.

It can be explain better through a scenario where young schoolchildren, meet in a lunch-break to have lunch together. The group is purely voluntary in nature and involves any character who wants to join the group. New student gets admission in the school and wants to have lunch with them. Therefore, new student will try to interact with the members of the existing group in order to get included in the group. After being included in the group, they will try to know each other and share their ideas and views. At some point, their discussion may lead to some conflict situations where the need for control arises which bring their focus back on lunch. The new student may get influence by the behaviour of some students in the group and wants to become friend with them. Thus, the need for affection becomes prominent. The model may help us in understanding the group activities better and may help us in resolving the problems of group behaviour.

4 RELATED WORK

There are researches going on that has a bearing in this field. The SGD Model (Synthetic Group Dynamics Model) [12] by Rui Prada is the prominent example. Experiments in SGD Model resulted in users increased trust and identification with the synthetic groups thus improved user experience. The model uses the dimensions of the five-factor model, which was given by J M Digman in 1990.

5 CONCLUSIONS & FUTURE WORK

This paper discusses the relevance of group dynamics in autonomous synthetic characters and throws light on the potential benefits that can be reaped by syncing group dynamics with various computational models of autonomous synthetic characters. It helps them to work in groups and behaves like human beings in a virtual world, which in turn helps us to deal with various situations.

Since the work is at initial stage therefore, the next step would be to work on building up a scenario that would help to demonstrate the group behaviour in synthetic characters and thus help them to attain the social believability. Later, we will find the mechanism for implementing the dimensions of the theory of group dynamics in autonomous synthetic characters.

REFERENCES

[1] Gustave Le Bon, The Crowd: A study of the popular mind, 1896.
[2] S. Freud, Group Psychology and the analysis of Ego, 1922.
[3] K. Lewin, (1935) A dynamic theory of personality. New York: McGraw-Hill.
[4] W. Schutz, The Interpersonal Underworld, Science & Behavior Books, 577 College Avenue, Palo Alto, California, 1966.

[5] R.S. Aylett and S. Louchart, (2007) Being there: Participants and Spectators in Interactive Narrative. M. Cavazza and S. Donikian (Eds.) ICVS 2007, LNCS 4871, pp 116-127, 2007. Springer-Verlag Berlin Heidelberg 2007.

[6] L. Hall, S. Woods, R. Aylett, L. Newall, and A. Paiva, (2005) Achieving empathic engagement through affective interaction with synthetic characters. Proceedings, International Conference on Affective Computing and Intelligent Interfaces, LNCS 3784, Springer, pp731-738

[7] A. Nazir, S. Enz, M.Y.Lim, R. Aylett, and A. Cawsey, Culture-Personality based Affective Model, Special Issue on Enculturing HCI in AI and Society Journal, Springer-Verlag London, 2009.

[8] M.Y. Lim and R. Aylett, An Emergent Emotion Model for An Affective Mobile Guide with Attitude, Applied Artificial Intelligence Journal, 23:835-854, 2009, Taylor & Francis Group

[9] M.Y. Lim, R. Aylett, J. Dias, A. Paiva, Creating Adaptive Affective Autonomous NPCs

[10] T. Narita and Y. Kitamura, Persuasive Conversational Agent with Persuasion Tactics, T. Ploug, P. Hasle, H. Oinas-Kukkonen (Eds.): PERSUASIVE 2010, LNCS 6137, pp. 15–26, 2010Springer-Verlag Berlin Heidelberg 2010

[11] I. Poggi, C. Pelachaud, F. de Rosis, V Carofiglio, B. De Carolis, Greta. A Believable Embodied Conversational Agent, IST project Magicster IST-1999-29078 with partners: University of Edinburgh, UK (coordination); DFKI, Germany; SICS, Sweden; Univ. of Bari, Italy; Univ. Of Rome, Italy; AvartarME, UK, T.H.E. Editor(s) (ed.), Book title, 1—6, yyyy Kluwer Academic Publishers.

[12] R. Prada, and A. Paiva, Teaming Up Humans with Autonomous Synthetic Characters, Artificial Intelligence, Elsevier, Vol. 173, No. 1, pg. 80-103, January,2009.

[13] R.Aylett and A. Paiva, Computational Modelling of Culture and Affect, Emotion Review, 2011.

[14] C. Bartl, D. Dörner, Comparing *the behavior of psi with human behaviour in the biolab game*. In F. E. Ritter, R. M. Young, eds.: Proceedings of the Second International Conference on Cognitive Modeling, Nottingham, Nottingham University Press (1998)

[15] D. Dörner, *The mathematics of emotions*. In Frank Detje, D.D., H. Schaub,., eds.: Proceedings of the Fifth International Conference on Cognitive Modeling, Bamberg, Germany (2003) 75–79

[16] D. Dörner, K. Hille, *Articial souls: Motivated emotional robots.* In: Proceedingsof the International Conference on Systems, Man and Cybernetics. (1995) 3828–3832

[17] G. Katsionis, M. Virvou, Adapting OCC Theory for affect perception in educational software,

Soil and Water conservation AdoPtion: the SWAP model - theory and policy applications using agent-based modeling

Peter George Johnson [1]

Abstract. Land degradation poses a threat to societies on a par with those of climate change and biodiversity loss. Soil erosion and degradation both play a large role in land degradation processes. Many measures and techniques are known to curb the effects of soil erosion and degradation, however adoption of these measures is often patchy or unsuccessful. An improved understanding of the adoption process, and new tools to aid stakeholder engagement and decision-making, are thus desirable. This paper reports on an ongoing project concerning the construction of an agent-based model (ABM) of the adoption process for soil and water conservation (SWC) measures amongst small-scale farmers (the SWAP model). The model implements a framework for farmers' behaviour developed in the literature and founded on a complex decision process that goes beyond simple rationality or utility-maximisation. The model aims to serve two purposes: first, to scrutinise the current theory on SWC; second, to explore potential policy interventions and aid decision-making and stakeholder engagement processes.The purpose of this paper is to present the project which can serve as a basis for discussion on various questions, presented in the conclusion.

1 INTRODUCTION

1.1 Land degradation and soil conservation

The UNEP states that land degradation poses a threat to the environment and society on a par with climate change and biodiversity loss [1]. SWC by farmers is a key part of the fight against land degradation and offers a way of helping deliver sustainable development to many parts of the world. Despite awareness of the problem, and the identification of simple SWC measures in many areas, the policy interventions designed to increase conservation adoption have often been a failure. Many writers have suggested this is because of poor calibration of policy to farmers and their behaviour (e.g., [2]). This is a result of the fact that land degradation is highly contextual [3].

1.2 SWC adoption

Farmers have been surveyed and interviewed in many areas across the globe. However, interventions have still struggled to regularly increase the level and efficacy of SWC adoption. It would appear that the extrapolation of individual household decisions to a wider community, and more broadly, constructions of farmer behaviour have been unsuccessful in some way. This is potentially due to the social and complex nature of the individual adoption decisions being made.

In developing countries investment-intense and technology-driven solutions to soil erosion and degradation have failed to gain widespread adoption for obvious reasons. However, it is less clear why relatively cheap, or traditional measures have not always been adopted successfully. Often farmers will adopt a measure, but only for a short time, or on a small area of land. Measures that require continued up-keep often fail after government or other extension services leave an area. Land tenure has also often been cited as key driver of the SWC decision (e.g., [4]). Farmers know they will not be on the land in the medium to long term, so decide to increase yields in the short run, rather than conserve the soil. Social dynamics can also play a key role in determining the success of a SWC measure; some measures may be socially unacceptable, or go against long standing norms.

Thus, we begin to get a feel of why the drivers of degradation, and SWC adoption, are strongly contextual. This poses a challenge to modelers looking to address the issue; an important question arises; how can we deal with and model a contextual problem, when trying to build models that can be applied to numerous cases?

1.3 Agent-based modeling

ABM offers an approach to the issue that has proven useful in the past when applied to this topic. Here, it allows for a qualitative understanding of the adoption decision of farmers (taken from the literature) to be implemented in a simulation, which then plays out these decisions in an iterative fashion. It is the ability to iterate this micro-dynamic that adds to the existing literature on the adoption process of SWC. Furthermore, the ability to initially include an exhaustive list of parameters in the model, and then attempt to reduce the model offers one potential solution to the question of how to build models for context dependent issues, as the smaller model can be quickly and easily applied to new cases.

1.4 Aim

1.4.1 The project

This project serves two purposes. First it allows us to analyse the qualitative decision rule developed in the theoretical literature. Second, once the model has been validated and refined, it can be used in two policy applications. Initially, to run hypothetical policy interventions under various scenarios. Primarily however, the model will serve as a potential tool to aid stakeholder discussion, engagement and decision-making on the ground. To this end, the model will be presented to various stakeholders at local and regional levels in

[1] Centre for Research in Social Simulation (CRESS), Univ. of Surrey, UK. Email: p.g.johnson@surrey.ac.uk

Ethiopia; stakeholders can critique the model, with the aim of increasing understanding around the potential use of the model as well as stakeholders evaluation of the model.

1.4.2 This paper

The aim of this paper is to present the project in its current position. The initial model has been built and is in the secondary development stage. The plan for finishing development and the main analysis has been designed. A workshop with stakeholders is currently being organised.

The topics that are probably of most interest to the symposia are: 1) the implementation and development of a decision behaviour developed in the existing literature to a high level of detail, and questions around the value in this method of deriving behaviour rules; and 2) the design and use of models for stakeholder engagement, discussion and decision-making. It is envisaged the model can be used to start a discussion around various questions, presented in the conclusion.

The paper continues as follows. Section 2 briefly overviews some relevant literature and details the decision framework adapted in the model. Section 3 presents the model in detail. Section 4 presents plans for model development and intended use. Section 5 concludes the paper, and raises discussion questions.

2 LITERATURE

2.1 Environmental management and ABM

There are a large number of studies applying ABM to various problems similar or relevant to that being studied here. These fall into a messy array of categories owing to the varied background and disciplines of the researchers addressing environmental management issues with ABM. The largest stream of literature concerns the modelling of land-use and land-cover change (LUCC). ABM has been utilised here because of its ability to incorporate spatial and ecological modeling alongside sophisticated modeling of human behaviour. [5] and [6] provide excellent, if slightly old, reviews which highlight the various approaches used.

Some recent studies have modelled farmer decisions explicitly, and linked them to ecological system models (e.g., [7] and [8]). However these often focus on the effect of behaviour on the environment, rather than the adoption of behaviours. Behaviour frameworks are often also based on economic analyses of behaviour alone. Others have focused on innovation diffusion (e.g., [9], [10] and [11]), highlighting the potential for ABM approaches in this area.

2.2 SWC adoption

There have been many studies considering the causes and factors associated with land degradation, soil erosion and degradation, and SWC measures (e.g., [12] and [13]). Surveys of households are often used alongside field data on soil condition and land management, to generate quantitative analyses. Qualitative studies have also been carried out focusing on farmers' perceptions and opinions on the causes of soil degradation and conservation adoption (e.g., [14]). This mature stream of literature has led to a reasonably well accepted understanding of how farmers decide on whether or not to adopt conservation measures. Though the decision and process of adoption is highly contextual, there is now an exhaustive list of potential factors identified in the literature.

Beyond this, a three-stage process (acceptance/information, adoption, and intensity/continued adoption) has been developed that is regularly put forward. [15] presents an explicit checklist of steps and factors that synthesises and crystallises the findings of the literature. It is this checklist, with associated factors that serves as the theory from which the agent behaviours in the ABM are directly derived.

2.3 The De Graaff et al (2008) decision framework

This section outlines the decision framework presented by Jan de Graaff and his co-authors [15], that is used in the ABM. The decision is split into three stages: the acceptance, adoption, and continued use stages.

2.3.1 Acceptance stage

Table 1. Steps and factors in the acceptance stage.

Steps	**Factors influencing acceptance**	**Agent attributes required**
1. Degradation symptoms recognised?	Perception of erosion problem, Off-farm employment.	Knowledge of land, Decision maker does labouring (Y/N).
2. Degradation effects recognised?	Age, Perception of erosion problem, Lack of education, Traditional beliefs.	Age, Knowledge of land, Education, Cultural inertia.
3. Degradation taken serious?	Their problem? Perception of erosion problem, Land tenure.	Adherence to social norms, Land tenure status,
4. Aware of conservation methods?	Lack of research/extension, Contacts with extension.	Knowledge of tech, Extension contact.
5. Able to undertake measures?	Labour availability, Age, Absence of farmer groups, Farm size, Income, Lack of credit, Land tenure.	Labour availability, Age, Social/group links, Number of fields, Income, Credit access, Land tenure status.
6. Willing to undertake measures?	Consumption requirement, Discount rate, Social status, Tribe, Gender, (Genuine) participation, Attitude, Family composition, Age, Off-farm income.	Consumption requirement, Discount rate, Cultural inertia, social links, Gender, Institution attitude, Successor? Age, Decision maker does labouring (Y/N).
7. Ready to undertake measures?	Few resources, Risk averse, Psychological threshold.	Income, savings, Risk attitude, Cultural inertia, social links.

Source: Adapted from [15]

2.3.2 The adoption stage

Once a positive decision to adopt has been taken, the intensity or effort of this adoption must be decided. Here the farmer decides how many of their fields to apply the adoption to. This is a function of the attributes of each field and the characteristics of the farmer (personal, economic, social and institutional links). The specific process is left unspecified.

2.3.3 The Continued-use stage

Once adoption has happened, the farmers still have the ongoing decision to continue using the conservation method, or to change their intensity/effort. The conditions that led them to adopt may change because of shifts in the agents characteristics, social and institutional links, or changes in the fields soil quality. This decision will be similar to the original adoption decision but with altered inputs in light of the fact the field/farm is currently under conservation methods.

3 THE MODEL

The model is presented using the ODD protocol developed in [16] and [17]. An ODD protocol is a tool used for the standardised description of ABM; primarily used to present models clearly, it is also a useful document for a modeler to produce early in the modeling cycle. ODD stands for overview, design concepts and details.

3.1 Overview

3.1.1 Purpose

The purpose of the ABM is to model the adoption decision of farmers for SWC measures. It is intended the model will help explore the current theory on adoption, on which the agents behaviour rule is based, and be applied to policy scenarios, as well as be developed as a stakeholder engagement, discussion and decision-making tool.

3.2 Entities, state variables and scales

The model contains agents that represent individual farming households (h/h) that have a decision between using two farming methods: non-SWC methods, or SWC methods. The households have a three-stage decision process in which they decide 1) whether they accept the need for SWC, 2) if they accept, how intensely they wish to adopt, and 3) once they have adopted, whether to continue adoption. These decisions (and the ability to carry out a behaviour once a decision has been made) are affected by the characteristics of the household and their local environment. The households interact with each other, influencing each others characteristics and thus decisions. The households decisions impact on their local environment, creating a feedback between the human and ecological systems in the model.

Farmer agents represent a farm household, and are by far the most complex agent, with several decision-making processes and multiple state variables (Table 2). The variables essentially represent the factors identified in the literature that affect adoption.

Table 2. Farmer agent state variables.

	Variable Name	Value	Notes
1	Position	Coordinates	Randomly distributed at initalisation
2	soil-conservation-decision	not accepted; accepted not adopted; adopted	Current status of decision
3	acceptance-decision-score	0-9	Current status of acceptance decision
4	age-of-decision-maker	Years	Norm dist
5	education-of-decision-maker	Years	Number of years of education of the decision maker in the h/h. Norm dist
6	Successor?	Y/N	Does the h/h head have a successor?
7	decision-maker-does-labouring	Y/N	Does the h/h head take part in farm labouring?
8	size-of-household	Persons	Norm dist
9	land-tenure-status	Owned / rented	
10	labour-access	Y/N	Does the h/h have access to hired labour?
11	credit-access	Y/N	Does the h/h have access to credit?
12	number-of-fields-owned	Number	
13	income	Number/Index	= number of fields owned * average soil quality of fields owned * knowledge of land
14	savings	Number/Index	Norm dist
15	consumption-requirement	Number/Index	= size of household * consumption-reqiurement-per-individual
16	risk-aversion	Score	Norm dist
17	discount-rate	Score	Norm dist
18	cultural-inertia	Score	Norm dist
19	adherence-to-norms	Score	Norm dist
20	institution- attitude	Score	Attitude towards outside institutions. Norm dist
21	influence-score	Score	Strength of influence on others. Norm dist
22	knowledge-of-land	Score	Norm dist
23	knowledge-of-technology	Score	Norm dist
24	extension-worker-contact	Y/N	Norm dist

N.B: Norm dist = Normally distributed around the case study data

Extension agents have only the most basic variable: position. After this they all have a fixed attribute for the distance they can move on each time step. Their only process is to pick a random heading and move forward in that direction on each time step. They function to effect farmer agents extension-worker-contact variable, when they are nearby.

The environment is modeled by many patches (number set by the user), which represent fields. A group of fields owned by a farmer agent makes up that agents farm. Each patch/field has the variables listed in Table 3.

Table 3. State variables of the environment/fields.

	Variable Name	Value	Notes
1	Position	Coordinates	
2	Soil-quality	Score	Norm Dist
3	Soil-conservation-practised?	Y/N	Is SC currently practiced on the field?
4	Owned-by	Agent ID	Shows farmer agent in charge of that field

There are state variables that do not belong to any specific agent, or are the same for all agents (globals), but can be changed at intialisation (see Table 4).

Shock weather events cause a sudden drop in soil quality and occur randomly. Farmer group vision determines how far farmers range of influence is, and over how large an area groups form. Death-age

Table 4. Global Variables.

	Variable Name	Value	Notes
1	Chance of shock weather event per tick	%	User defined.
2	Farmer group vision	Score	User defined.
3	Death-age	Years	User defined.

indicates the age of death for a h/h decision maker. When agents die, agents with successors will sprout new agents with similar characteristics, agents without successors will sprout new agents with random characteristics.

Each time step represents three months. The model can be run for 10, 25, 50 or infinite year lengths (ie:, 40 ticks, 100 ticks or 200 ticks). This time scale is somewhat arbitrary and reflects a rough approximation of how long the relevant decisions take to make and implement in real life, as well as a consideration of relevant policy decision time frames.

The patches do not represent an explicit size, but rather a non-size specific field. Farmer agents may own different numbers of fields, this is randomly generated at the initialisation. The average size of farms could be set at the initialisation if it was felt this was parameter worth exploring.

3.2.1 Process overview and scheduling.

The basic processes of the model involve the decision of farmer agents to adopt SWC measures. This is done in three parts, first an acceptance decision must be reached. Second an adoption decision must be made, before a continued use decision can be made.

The first two UML diagrams (Figures 1 and 2) detail this high-level process description. Figure 1 shows the three different decisions that an agent must choose between on each time step. First they must check their current decision scores and choose the appropriate decision to make this time step.

If the farmer agent is still in the acceptance decision (Figure 2), they will again check their current decision, and choose which step is next for them to consider. Here, only one step in the decision process can be made in each time step.

Within the acceptance decision (Figure 2), the eight steps are made using the following (quasi code):

```
1) Run symptoms recognised
if   ( farm soil quality = low )
and  ( decision maker works on farm )
and  ( farmer knows the land well )
then  [ recognise symptoms ]

2) Run effects recognised
if   ( farmer not too old )
and  ( farmer knows the land well )
and  ( farmer is well educated )
and  ( farmer has extension contact )
and  ( farmer has low culturalinertia )
then  [ recognise effects ]

3) Run degradation taken seriously
if   ( farmer has extension contact )
and  ( farmer owns the land )
then  [ take degradation seriously ]

4) Run aware of SWC methods
if   ( farmer has knowledge of methods )
and  ( farmer has extension contact )
then  [ be aware of methods ]

5) Run able to undertake SWC
if   ( farmer can hire labour )
and  ( farmer not too old )
and  ( farmer has extension contact )
and  ( farmer can access credit )
and  ( farmer owns the land )
then [ able to undertake SWC ]

6) Run willing to undertake SWC
if   ( discount rate is low )
and  ( farmer has low cultural inertia )
and  ( farmer  sympathetic to gov/NGOs )
and  ( farmer has a family successor )
and  ( farmer is not too old )
and  ( decision maker works on farm )
then  [ willing to undertake SWC ]

7) Run ready to undertake SWC
if   ( not too risk averse )
and  ( farmer has enough savings )
and  ( farmer has enough income )
then  [ ready to undertake SWC ]

8) Run accept SWC
[ set acceptance score to:
accepted but not adopted ]
```

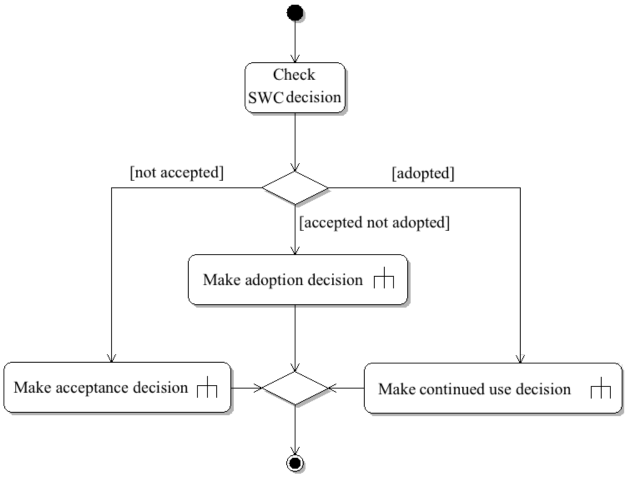


Figure 1. Agent's basic decisions.

Note, at each time-step there is 5% chance that the agent will jump to the next decision point, this represents an element of chance or noise in the decisions.

If the agent is in the adoption decision process, they must decide on how much of their farm they want to adopt SWC. The amount of land they adopt conservation on is determined by their level of savings (savings must meet a minimum threshold), their contact with extension workers (contact is required for any adoption), and their risk aversion score (less risk averse agents will adopt a higher level).

Finally, if they have already adopted SWC measures, they must decide whether to increase or decrease adoption. If their income is higher than their consumption requirement they will increase adoption by 20%. If their income is lower than their consumption requirement they will reduce their adoption by 20% (Note: the presence of adoption will increase soil quality, which in turn will increase income). Consumption rate is set by the user, and income is a function of the soil quality and farmer knowledge.

Figure 3 shows the basic processes behind the interaction of agents with each other.

The following quasi-code details the changes that are made to agent state variables on each time step.

```
Farmers

Ask farmers

[ increase age 0.25years ] and
[ increase knowledge of land 0.25 ] and
[ recalculate current farm soil quality ] and
[ recalculate current income ] and
[ check for extension agents nearby ] and
[ die and spawn successor? ]

The environment

Ask fields

[ shock weather event?
If yes [ reduce soil quality of all  patches ] ]
and  [ random change in soil quality ]
and  [ if soil conservation present
then [ improve soil quality ] ]
and  [ if fields nearby have good soil quality
then [ increase mine ] ]
and  [ if fields nearby have poor soil  quality
then [ decrease mine ] ]
```

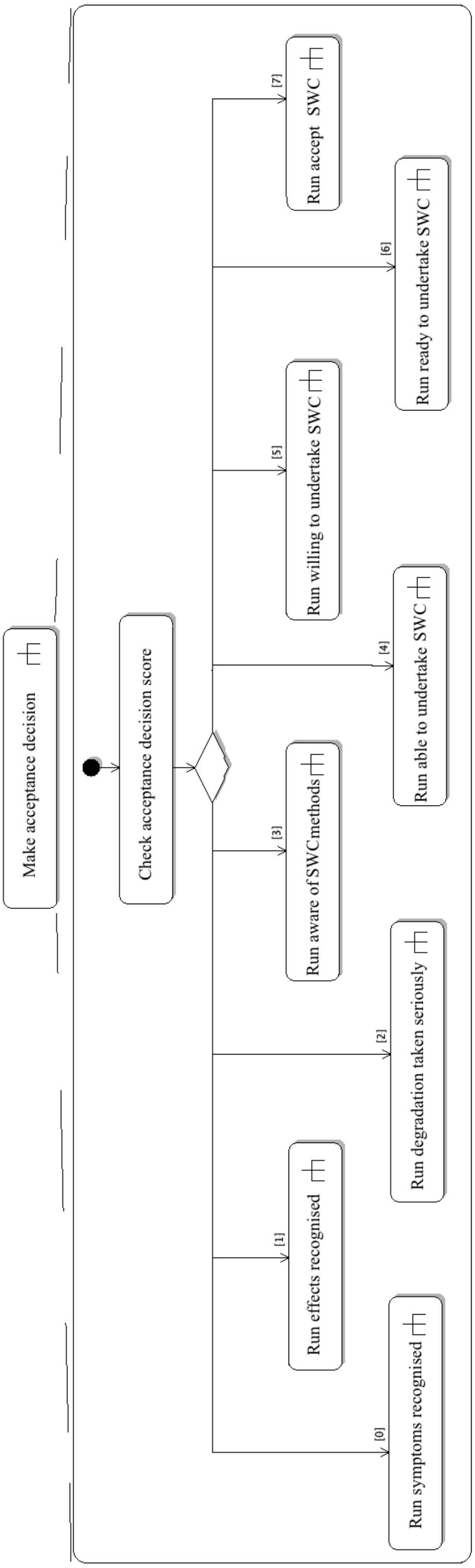

Make acceptance decision
Check acceptance decision score
[0]
Run symptoms recognised
[1]
Run effects recognised
[2]
Run degradation taken seriously
[3]
Run aware of SWCmethods
[4]
Run able to undertake SWC
[5]
Run willing to undertake SWC
[6]
Run ready to undertake SWC
[7]
Run accept SWC

Figure 2. Agent's acceptance decision.

The processes are carried out by the agents one at a time, but in a randomised order each time step.

In one time-step an agent can pick a decision, and carry it out, but only one decision, when they change their decision score they must stop for that time-step (i.e., an agent can decide they recognise the existence of land degradation, but cant then also suddenly be aware of methods to combat it; or an agent can decide they do accept the need for SWC, but then cant also decide how much to adopt).

3.3 Design concepts basic principles

3.3.1 Basic principles

The model of adoption behaviour of farmers is based on a framework explicitly stated in [15]. Though the issue is highly contextual and factors are different in every case, a comprehensive list of all factors that affect adoption is reasonably settled in the literature, with many studies covering similar social, economic, historical, political and other anthropogenic factors.

On top of this framework, the model incorporates interaction between the farmers, affecting each others variables; the framework for this is novel.

3.3.2 Emergence

The spatial pattern of adoption is emergent. All the other key macro variables of interest are not technically emergent.

3.3.3 Adaptation

The agents do not adapt their decision process.

3.3.4 Objectives

The agents have no explicit objective when making decisions, they simply make a decision if they fulfil all the criteria necessary, (i.e., it is assumed that SWC will increase utility for the agents, and that they inherently know this, they just have to get to the point where they can accept the need for SWC, and are able to adopt it).

3.3.5 Learning

The agents do not learn.

3.3.6 Prediction

Agents use their current situation to make predictions and thus decisions for the next time period.

3.3.7 Sensing

The agents can sense the soil quality of their fields, and thus whole farm. They can sense the attributes of other agents when interacting (see figure 3).

3.3.8 Interaction

See figure 3 for a full UML description of agent interaction. This framework was developed for this model and is not based on any specific previous literature; rather, it is based on the understanding of common ways in which farmers interact (e.g., through trade unions, through local leaders).

3.3.9 Stochasticity

Shock weather events, which reduce soil quality significantly, are modeled stochastically.

3.3.10 Collectives

Some simple collectives are modeled when agents interact to influence each other. This may be in the form of a group of geographically close agents, whom either all influence each other, or, follow the influence of a leader.

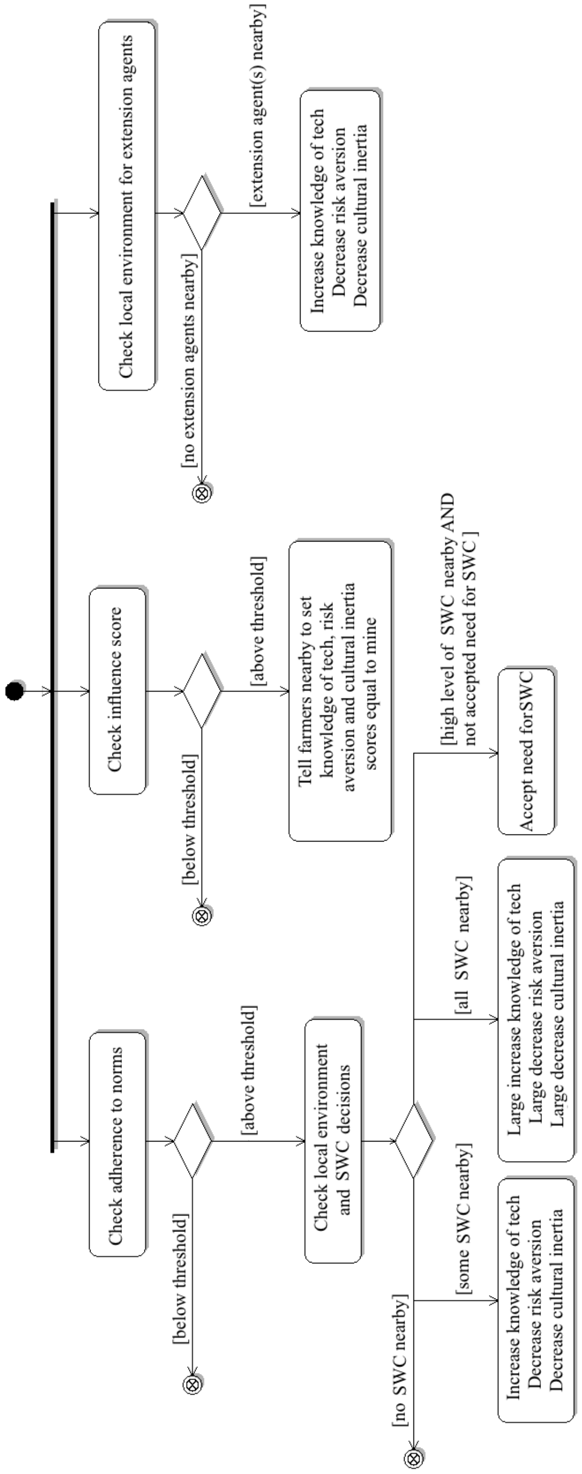

Check local environment for extension agents
[extension agent(s) nearby]
Increase knowledge of tech
Decrease risk aversion
Decrease cultural inertia
[no extension agents nearby]
Check influence score
[above threshold]
Tell farmers nearby to set knowledge of tech, risk aversion and cultural inertia scores equal to mine
[below threshold]
Check adherence to norms
[above threshold]
Check local environment and SWC decisions
[below threshold]
[high level of SWC nearby AND not accepted need for SWC]
Accept need forSWC
[all SWC nearby]
Large increase knowledge of tech
Large decrease risk aversion
Large decrease cultural inertia
[some SWC nearby]
Increase knowledge of tech
Decrease risk aversion
Decrease cultural inertia
[no SWC nearby]

Figure 3. Agents interaction.

3.3.11 Observation

The main visualization window is observed for initial qualitative assessment. The adoption rates of farmers, and the rate of adoption on fields are recorded.

3.4 Details

3.4.1 Initialisation

The initalisation of parameters is determined by the input data from each case study. All variables that do not have available data, are set at plausible levels and, if farmer agent variables, are normally distributed across the population of agents.

3.4.2 Input data

Data is derived from census data, previous studies and other secondary data sources.

3.4.3 Submodels

All the details of the model are included above.

4 MODEL DEVELOPMENT AND USES

4.1 Model development

The model development is on-going at the time of writing. Once the theory had been implemented in the model in its most basic form, data was gathered on one case study area Tigray in northern Ethiopia. The model was then put through a developmental sensitivity analysis (SA). This was done with two key purposes: firstly as a continuation of the verification and bug checking process, and secondly, and primarily, as part of a model development stage aimed at reducing the number of parameters in the model. The number of parameters from the theory was relatively high for this type of ABM, therefore it was important to see if the model could be made more simple without reducing its explanatory power. This reduction would thus be a potential addition to the theory on conservation adoption, made possible by the use of ABM to iterate the micro-dynamics already identified in the literature.

The SA comprises two stages, first two local SA were conducted, with the various farmer agent parameters having sensitivity scores calculated and a stepwise regression being used to find which parameters had the most power in explaining the output of the model the level of adoption of SWC. The next stage, which is yet to be conducted, is to run an interaction SA on a subset of parameters. These parameters are chosen on the basis of what parameters were most powerful in explaining the output (theoretically interesting), and which parameters are most amenable to policy interventions (policy relevant).

Once this interaction SA has been conducted the model will be reduced in size in two ways, once to include only the most theoretically interesting parameters those that were most powerful in explaining the output, and second to include only those parameters that are amenable to policy interventions. Then we will be left with three models, the comprehensive model, the theory model and the policy model.

4.2 Intended uses

4.2.1 Theory development

The performance of the comprehensive and theory model will be analysed for potential insights into the existing theory on SWC adoption. Real world data will be used to validate the models on several case studies. The two envisaged outcomes will be the supporting or undermining of the current theory, and the identification of a potentially more parsimonious model/theory.

4.2.2 Policy applications

The comprehensive and policy models, once validated successfully, will be used to run hypothetical policy scenarios. Though interesting, these scenarios on their own are unlikely to have any real policy value; they would represent outputs from a model policy-makers and stakeholders would be unlikely to trust.

To address this, the model will go through a further stage of development, with the direct input of stakeholders and decision makers. This will be in the form of workshops run with leaders and experts at the local and regional level in Ethiopia. The purpose of the workshops will be to present the model(s), receive feedback on the models directly, and improve understanding around what form models like this can take that is most useful for stakeholders. The workshop would also allow for development in the model's final application, as a stakeholder engagement, discussion and decision- making tool.

5 CONCLUSIONS AND DISCUSSION QUESTIONS

This paper has presented the ongoing development of the SWAP model of SWC adoption amongst small-scale farmers. The model behaviour rules are based on a decision framework taken from the existing literature on SWC adoption. The model is intended to serve as a tool for an analysis of the existing theory, to explore patterns when it is iterated, and consider its validity. Beyond this, the model has potential policy applications. Policy scenarios can be run on a successfully validated version of the model. However, arguably most important is the potential for the model to serve as a tool of stakeholder engagement and decision-making. To explore this the model is being presented to stakeholders in June 2013.

It is hoped this project will improve understanding and add to the literature on the use of ABM as a stakeholder engagement tool and aid to decision-making. The models power lies not in pure forecasting, but rather in aiding stakeholder communication and understanding.

In the context of the SOCIAL.PATH symposia, the following questions are of interest to the author:

1) What are the pros and cons of using a highly developed theory of individual behaviour directly in an ABM?

a. Is this a valid way of deriving behaviour rules?

b. Could this help computer scientists bypass the need for advanced social science skills/knowledge?

2) How can modellers deal with context dependent issues?

a. What can we do to make a model easy to apply to new cases?

3) What common issues arise when using a model for stakeholder engagement, discussion and decision-making?

a. How can modellers deal with the tension between stakeholders wanting a simple and clear model, and the desire for realism and detail in a model?

REFERENCES

[1] United Nations Environment Programme, *Global Environment Outlook 4*, (2007)

[2] P. Illukpitiya and C. Gopalakrishnan, 'Decision-making in soil conservation: application of a behavioral model to potato farmers in Sri Lanka', *Land Use Policy*, **21(4)**, 321-331, (2004).

[3] A. Warren, 'Land degradation is contextual', *Land Degradation & Development*, **13(6)**, 449-459, (2002).

[4] T. Niazi, 'Land Tenure, Land Use, and Land Degradation: A Case for Sustainable Development in Pakistan', *The Journal of Environment & Development*, **12(3)**, 275-294. (2003).

[5] R.B. Matthews, N. Gilbert, A. Roach, G. Polhill and N. Gotts, 'Agent-based land-use models: a review of applications', *Landscape Ecology*, **22(10)**, 1447-1459, (2007).

[6] D. Parker, S. Manson, M. Janssen, M. Hoffmann and P. Deadman, 'Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review', *Annals of the Association of American Geographers*, **93(2)**, 314-337, (2003).

[7] T. L. Ng, J. W. Eheart, X. Cai, and J. B. Braden, 'An agent-based model of farmer decision-making and water quality impacts at the watershed scale under markets for carbon allowances and a second-generation biofuel crop', *Water Resources Research*, **47**, W09519, (2011).

[8] D. Valbuena, P.H. Verburg, A. Veldkamp, A.K. Bregt and A. Ligtenberg, 'Effects of farmers' decisions on the landscape structure of a Dutch rural region: An agent-based approach', *Landscape and Urban Planning*, **97(2)**, 98-110, (2010).

[9] T. Berger, 'Agent-based spatial models applied to agriculture : a simulation tool for technology diffusion , resource use changes and policy analysis', *Agricultural Economics*, **25(0169)**, 245-260, (2001).

[10] P. Schreinemachers, T. Berger, Sirijinda, A and S. Praneetvatakul, 'The Diffusion of Greenhouse Agriculture in Northern Thailand: Combining Econometrics and Agent-Based Modeling', *Canadian Journal of Agricultural Economics*, **57(4)**, 513-536, (2009).

[11] P. Schreinemachers, C. Potchanasin, T. Berger, and S. Roygrong, 'Agent-based modeling for ex ante assessment of tree crop innovations: litchis in northern Thailand', *Agricultural Economics*, **41(6)**, 519-536, (2010).

[12] R. Gauthier, 'Agro-ecological strategies in North Lampung, Indonesia: social constraints to biological management of soil fertility.' *Netherlands journal of agricultural science*, **48(1)**, 91-104, (2000).

[13] M.A. Damisa, and E. Igonoh, 'An Evaluation of the Adoption of Integrated Soil Fertility Management Practices among Women Farmers in Danja, Nigeria', *The Journal of Agricultural Education and Extension*, **13(2)**, 107-116, (2007).

[14] K.B. Wilson, 'Water Used to be Scattered in the Landscape: Local Understandings of Soil Erosion and Land Use Planning in Southern Zimbabwe', *Environment and History*, **1(3)**, 281-296, (1995).

[15] J. de Graaff, A. Amsalu, F. Bodnar, A. Kessler, H. Posthumus and A. Tenge, 'Factors influencing adoption and continued use of long-term soil and water conservation measures in five developing countries', *Applied Geography*, **28(4)**, 271-280, (2008).

[16] V. Grimm et al, 'A standard protocol for describing individual-based and agent-based models', *Ecological Modelling*, **198(1-2)**, 115-126, (2006).

[17] V. Grimm, U. Berger, D.L. DeAngelis, G. Polhill, J. Giske and S. Railsback, 'The ODD protocol: A review and first update', *Ecological Modelling*, **221(23)**, 2760-2768, (2010).

Regulism, regularism and some limitations of agent-based modelling

Rodger Kibble[1]

Abstract. The emerging field of Normative Multi-Agent Systems has the twin goals of providing tools for simulating human societies and coordinating activities among heterogenous autonomous software agents in an open environment. Norms are frequently modelled as either observable regularities or explicit precepts. We consider arguments from analytic philosophy that neither approach can account for the origins of norms as both are prone to regress problems, and we note that a sample of simulation models all assume some form of built-in normativity. We conclude by sketching some implications for agent design.

1 INTRODUCTION

While this paper was in preparation, the British House of Commons voted in favour of legalising marriage between couples of the same sex. This came just 10 years after the repeal of the so-called “Section 28” which had banned schoolteachers from presenting homosexuality as a “pretended family relationship”. In the run-up to the parliamentary vote, an opinion poll showed that three out of five voters supported the measure. As recently as 1967, homosexual relations between men in the UK were classed as a criminal offence, as they still are in many countries world-wide. Clearly there have been some quite fundamental shifts in social attitudes in the last few decades, which have if anything picked up speed in recent years: the idea that same-sex marriage could become an accepted part of social life in the UK would have been almost unimaginable just a generation ago. This shift is by no means complete as there is still a substantial minority opposed to equal marriage, particularly among the leaders of organised religious groups, and it is quite conceivable that there will never be 100% acceptance.

One of the aims of the emerging field of normative multi-agent systems is to seek to understand these kinds of phenomena better by modelling societies with collections of artificial software agents and seeing how agents can be made to adopt, propagate and act on normative beliefs, and how norms can spread within a society. Research on normative MAS has the complementary purpose of developing techniques for restricting the autonomy of artificial agents in order that they may cooperate in a productive way. Although these two research programmes have distinct goals and assumptions there are numerous areas of overlap in their subject matter and insights from one field may well inform the other. This paper takes a critical look at some developments in this field from the perspective of analytical philosophy and poses the question of whether agent-based models are capable in principle of simulating the emergence and recognition of norms. The paper extends and deepens the analysis of [14], considering a wider range of examples and theoretical issues.

1.1 SOCIAL NORMS

As is often observed, the term “norm” is hard to pin down to a precise definition and is used in various difference ways in the literature on normative multi-agent systems [12]. This can make it difficult to compare the results of different research projects. Some uses of the term include:

1. An observable regularity in a society. This makes no assumptions about the agents’ internal architecture or their mental capacities, if any.
2. Conditioned behaviour: a regularity that is reinforced by punishment or reward.
3. A convention in the sense of [15]: a joint solution to a coordination problem which has multiple equlibria.
4. A pattern of behaviour which is mutually expected, characterised by mutual accountability among actors.
5. A legal norm, which is either enacted by a legislative body or derived by judges from custom and precedent.

A norm can be breached in various ways: if the norm is prescriptive, it is breached by acting in a non-approved manner; if it is permissive, it is breached by trying to stop people acting in accord with it. Some researchers have equated norms with conventions as characterised above, but I think it is helpful to maintain a distinction and will assume the stronger definitions (4) or (5) when referring to norms. Agents may have purely instrumental reasons to comply with conventions as described by Lewis, while it is questionable whether conformance with norms can be reduced to instrumental considerations. Moreover, norms can lead agents to act contrary to their own interests [6] – for example, [1] opens by relating how in 1804, Alexander Hamilton felt obliged by a social norm of “honour” to fight a duel which resulted in his death. Norms have also been characterised in a purely behavioural way, without overt reference to the attitudes of participants:

> DEFINITION. A norm exists in a given social setting to the extent that individuals usually act in a certain way and are often punished when seen not to be acting in this way. (Axelrod [1: 1097])

This seems to fall midway between conventions and norms proper: while agents can be motivated to “act in a certain way” simply in order to avoid punishment, it is harder to explain in instrumental terms why other agents would take on the cost of punishing.

The “strong” version of norms I have hinted at above involves the notion of *mutual accountability* between members of a society. That is, agents are liable to sanctions if they violate accepted norms, but those administering sanctions must also be ready to account for their actions. Agents are therefore assumed to have communicative and argumentative competence, and

[1] Dept of Computing, Goldsmiths College, London SE14 6NW. Email: r.kibble@gold.ac.uk

norms emerge and are sustained through deliberation and negotiation. This tradition is associated with the work of Brandom [3,4], Habermas [8], Heath [9,10], Rouse [18] and others, and is generally less well-known within the MAS community.

1.2 NORMATIVE MAS

Approaches to normative MAS have been classified in various ways by different authors according to the particular problem areas under discussion and whether the focus is more on engineering principles for coordinating artificial societies or on simulating actual human society. A recent survey of the state of the art in normative multi-agent systems [12] proposes a "consensus" model of the "norm life cycle" incorporating the processes of creation, transmission, recognition, enforcement, acceptance, modification, internalisation, emergence, forgetting and evolution. They distinguish in terms of *norm origins* between *Type I* norms, which are decreed by an authority, and *Type II* which emerge from interactions between agents. For instance, agents may exist within an *electronic institution* which monitors and regulates agent interactions according to rules which may themselves adapt and evolve in response to agent behaviour (op cit 4.7). They note three methods of norm creation in the "natural world": decree by an agent in power, spontaneous emergence, and negotiation by agents within a group. Note that a Type II norm may become sufficiently accepted that it ends up being codified as a Type I decree or statute: for example [1] notes increasing intolerance towards public smoking and presciently speculates that this may lead to the enactment of anti-smoking laws (op cit: 1106).

Neumann [17] compares simulation models in terms of two different distinctions: on the one hand the social-theoretic contrast between methodological individualism and role-based accounts such as those of Durkheim and Parsons, and on the other the implementation technologies inspired respectively by game theory (stemming from Axelrod's classic 1986 paper) and AI/cognitive models. The studies in the game-theoretic tradition generally deal with Type II, emergent norms while those with a AI-based or cognitive tradition deal with both Types I and II. He states that there has been a "paradigm shift" in the last 20 years towards methodological individualism, in reaction to the dominance of classic role- and value-based theories which were seen to simply treat social structures as "given" without attempting to account for their origins. The game-theoretic agent-based approach to social simulation is claimed to be aligned with this programme: "in agent-based simulation models (Artificial Societies), structures emerge from individual interaction". The challenge for models inspired by methodological individualism is to account for the emergence of institutions and shared values on the basis of individual preferences and instrumental reasoning, without positing an initial framework of rules governing interaction. Hodgson [11] claims on the basis of a survey of economic theory and analysis that "The narrow methodological individualist has a problem of infinite regress: attempts to explain each emergent layer of institutions always rely on previous institutions and rules." We will examine in this paper whether a similar regress arises for ABM.

This paper looks at normative agent models from yet another perspective, that of analytic philosophy. There is a long-standing critique within this literature of approaches to normativity which [3] dubs *regulism* and *regularism*, which essentially treat norms as either explicit precepts or statistical regularities, possibly enforced by sanctions in either case. There are at least superficial similarities here with Hollander and Wu's Type I/II distinction. The next section of this paper will begin by setting out the critique and considering whether it actually applies in principle to agent-based simulations of normative systems, and will proceed to look at selected models in more detail.

2. REGULISM AND REGULARISM

Regulism may be seen as a generalisation of Type I and construes norms as rules or precepts, which may for example be laid down and enforced by some authority or explicitly agreed among agents by means of a contract or treaty [3:18ff]. Regularism corresponds to a "behaviourist" variant of Type II, according to which norms are quantifiable regularities in the behaviour of members of a community (op cit: 26ff). Axelrod's definition above can be seen as a generalisation of regularism, where both the norms themselves and positive or negative sanctions which reinforce them are specified as probabilistic regularities. Brandom argues that both these notions are essentially incoherent and prone to regress, for reasons which are explained in the remainder of this section.

2.1 REGULISM AND TYPE I NORMS

It might seem that Type II norms would be weaker than Type I as they are not backed up by any authority or contract, but rely on the tacit agreement of all members of a community. In fact, Brandom's analysis implies that the reverse is true as Type I norms depend on Type II to have any force. The flaw in regulism is that agents need to be subject to not only the rules that constitute explicit norms, but rules that tell them how to follow a rule – and indeed, norms of obeying the edicts of a particular authority or of honouring contracts. This, it is argued, gives rise to a regress which must eventually be grounded in rules that are implicit in practice. In human societies, legislatures and law enforcement agencies rely on a general disposition among the populace to conform to expected standards of behaviour or what Heath [9: 155] calls a "norm-conformative orientation". There seems further to be an explanatory gap between the spread of normative beliefs and the adoption of norm-governed behaviour, perhaps best encapsulated in the classic "free rider" problem. To recognise that a normative belief exists is not the same as adopting it; and adopting a normative belief may not be sufficient to consider oneself bound by it.

In an implemented multi-agent system the grounding would ultimately be provided by *design decisions* which determine what rules and procedures are either implicit in the agents' practice or explicitly coded in their internal architecture. This could translate to *offline design* of the Type I norm itself or of a disposition to obey edicts from a particular authority; this means that the designer would decide "what norms a system will follow and encode them directly into the agents" [12]. Hollander and Wu note that in the majority of normative systems, norms are either "designed off-line and implicitly part of the agent's behaviour" or are "explicitly represented in the agents".

I suggest the regulist approach is also vulnerable to another kind of regress. Whatever authority is responsible for decreeing and enforcing the norms must consist of a group or class rather than a single individual: no one agent or Hobbesian Sovereign can be constantly monitoring the actions of every member of a community, in any realistic setup. (Even Stalin or Saddam had to sleep.) But then this governing class must itself act with a common purpose, following norms that pertain within the group; and so the problem of order re-emerges within the "authority". This objection applies to human societies and to MAS which are explicitly intended to model emergence of natural norms; in a purely artificial society it is of course plausible that there may be a single, persistent governing module which is able to monitor and direct the actions of other members of the society.

2.2 REGULARISM AND TYPE II NORMS

Turning to emergent Type II norms, agents' behaviour is also ultimately constrained by the basic preferences specified by the designer and the way their environment has been set up. For example, [7] argue that the results obtained by Axelrod [1], apparently modelling the emergence of a stable norm of cooperation, "are dependent on very specific and arbitrary conditions without which the conclusions tend to change significantly". We return to this issue in section 3.1.

Brandom and Rouse accuse regularists of what Brandom calls "gerrymandering" [3: 28-9]: the claim is that there is no uniquely identifiable sequence of actions that make up a norm-conformative performance. In principle, a sequence of events can be seen to instantiate any number of putative regularities, and fixing on one of them as the "norm" which is supposed to be at play is in itself to make a normative judgment. The term "gerrymandering" is perhaps unfortunate as it carries a connotation of wilful manipulation or deceit; I will continue to use it as a term of art, but retaining scare quotes in order to disown this connotation. Winch [26: 29] notes in his discussion of Wittgenstein's account of rule-following that "any series of actions which a man [sic] may perform can be brought within the scope of some formula or other if we are prepared to make it sufficiently complicated" and as part of a critique of Weber, further questions the notion that statistical analysis can provide a sociological explanation of patterns of behaviour: "The compatibility of an interpretation with the statistics does not prove its validity … one might be able to make predictions of great accuracy … and still not understand what those people were doing" (op cit: 113-115).

To be honest, the "gerrymandering" argument seems excessively sceptical. In principle it is no doubt true that one can devise a multitude of descriptions for a particular series of events given sufficient ingenuity, but it seems reasonable to assume that members of an agent society are able to discriminate different types of action and to perceive some as more relevant than others to their immediate purposes. And while a given sequence of events may be interpretable as instantiating any number of possible regularities, only a limited subset of these classifications will generally prove useful for predicting whether future events fall into the same class. Taken to an extreme, this argument would imply that we can never learn concepts, as any given set of instances would have an indefinite variety of properties in common. Having said this, it is indubitable that many agent-based simulations leave themselves open to an accusation of "gerrymandering" in that they generally assume an extremely parsimonious ontology, such that agents only discriminate between a very small set of action-types. In Axelrod's classic experiments the repertoire is limited to choices of whether or not to *defect* or *punish*, while the more recent [20] have two classes of agent, one of which can *arrive*, *eat*, *pay*, *tip* and *depart*, while the other can *wait* [at table] or *sanction*. The question of how these categories are extracted from the rich and varied patterns of everyday behaviour is disregarded.

Regularism in the form in which it is articulated by Axelrod also runs into a regress problem since sanctioning is itself a norm-governed activity which may be done properly or not; someone who wrongly sanctions an action may themselves be properly subject to sanctions either by their sanctionee or some public-spirited third party (see [3: 42-46] on *normative sanctions*). For example, an apparent deviation from a norm may itself have been intended to sanction another agent's deviance. This argument is consistent with the notion of a normative social practice found in [18], which is "maintained by interactions among its constitutive performances that express their mutual accountability. Such holding to account is itself integral to the practice and can likewise be done correctly or incorrectly". Rouse (op cit) claims that the cycle of holding performances to account, holding those holding-to-accounts to account and so on "need never terminate in an objectively characterizable social regularity".

Furthermore, what counts as a sanction is itself normatively determined and often purely symbolic in nature: as Heath [9: 154] observes, "most social sanctions do not have any intrinsic punitive quality". So for example if someone breaches certain rules of the road, other drivers may shout or gesture at him or sound their horns: none of these actions results in any actual harm to the individual. Or if I suspect someone of telling lies (violating a norm of truth-telling), I might say no more than "Are you sure about that?" or "That's rather an unusual story, isn't it?" [16]. Thus one cannot simply identify a norm by looking out for behaviour which is often "punished" without first having an understanding of what constitutes "punishment".

2.3 SUMMARY

To summarise the position we have arrived at so far:

1. Norms cannot in general be identified with explicit rules, since the efficacy of these rules relies on "rules for following rules" which must ultimately be implicit in social practice.
2. Norms cannot be identified with conventions in the sense of [15], since the latter but not the former can generally be understood in purely instrumental terms.
3. A norm cannot exist on its own. There must be some way of indicating whether or not someone has followed a norm correctly by means of "sanctions"; and what counts as a sanction, and whether or not it has been correctly performed, are both normatively constituted.
4. Sanctions are not one-way actions but are reciprocal: someone who has been sanctioned may challenge the action, claiming that it was unjustified, in effect sanctioning the sanctioner.

It looks worryingly as if we are being pushed towards a *holistic* view of norms. Just as Sellars and Brandom maintain that "one must have many concepts in order to have any" [3:89] it may be that one must have a grasp of many norms in order to operate with any. None of this is to deny that explicit precepts or probabilities have a place in the analysis of normativity. The argument that norms must ultimately be grounded in social practice does not entail that all norms are immediately so grounded.

3. MODELS

This section considers a sample of simulation models in the light of the above discussion: firstly we briefly review Axelrod [1] and the critical re-evaluation of his contribution by [7], then we will examine one model of *norm emergence* [22] and three of *norm recognition* [22, 5, 2].

3.1 AXELROD (1986): AN EVOLUTIONARY APPROACH TO NORMS.

Axelrod [1] is often cited as a classic point of reference in the game-theoretic approach to modelling norms, yet his evolutionary simulation actually takes up a relatively small part of the paper. His proposal is not linked in any systematic way to real-world scenarios, though he suggests "cheating in an exam" as an example of *defecting*. Axelrod informally describes eight proposed mechanisms for supporting norms including metanorms, reputation, deterrence and law, while only the first of these is given a computational model and the possibility that metanorms do actually play a part in sustaining norms "remains speculative" (op cit: 1103).

Briefly, Axelrod sets up a regime according to which any agent is not only expected to punish defectors, but is also entitled to punish those who fail to administer punishment (the "metanorms" model). This is reminiscent of the principle which [9] attributes to Durkheim, that a norm-conformant agent is disposed not only to punish deviants but to punish those who do not sanction deviance.

Axelrod found that a stable cooperation state emerged only when players were initialised with a high level of *vengefulness*, i.e. a high probability "that the player will punish someone who is defecting" [1: 1098]. Crucially in Axelrod's model, the cost to the punisher and the defector are the same for enforcement of both the cooperation norm and the metanorm, namely -2 and -9 respectively and players are equally vengeful at both levels. Galan and Izquierdo [7] considered this to be unrealistic and reran the simulation with the meta-enforcement cost and meta-punishment payoff divided by 10 (-0.2 and -0.9). The result was that the cooperation norm quickly collapsed and this state was "sustained in the long term". The authors draw the conclusion that the original results were highly dependent on the initial settings, and argue that it is essential for any agent-based simulation to be replicated by independent researchers - in part to uncover any assumptions which the original developer may have been unaware of or considered to be unimportant.

Axelrod's notion of metanorms has other shortcomings. As noted above, agents have the ability to punish those who fail to punish defections, but the insights of Brandom and Rouse would require that a metanorm is treated as a fully-fledged norm: agents should also have the capacity to sanction both failure to enforce a metanorm (and so on recursively), and punishment that is wrongly administered at any level of normative enforcement

It is doubtful whether the evolutionary model was intended to provide a complete account of norm emergence and stabilization, just as it seems unlikely that many of the historical examples of norms and discussed in the paper are actually sustained by "meta-punishment": e.g. tolerance of political opposition (p. 1095), aversion to the use of chemical or nuclear weapons (p. 1096), or the declining acceptance of "the right to smoke in public without asking permission" (ibid.). Rather, the stability of these kinds of norms or their replacement by new norms surely involves explicit deliberation, negotiation and argumentation in the public sphere – none of which is modelled in Axelrod's system. Neumann [17] points out that while agents in the metanorms model may appear to an observer to be conforming to a norm, in fact "agents do not act because they want to obey (or deviate from) a norm. They do not `know' norms".

3.2 SEN AND AIRIAU (2007): EMERGENCE OF NORMS THROUGH SOCIAL LEARNING.

Sen and Airiau [22] treat norm emergence as a problem of resolving social dilemmas where there are multiple game-theoretic equilibria. Thus they tacitly equate norms with *conventions* in the sense of [15]. The particular scenario investigated is the emergence of "rules of the road", with particular sub-problems of whether to drive on the left or the right and who should yield at a junction. The setup is that at each iteration, every agent is randomly paired with a randomly selected agent to play a social dilemma game. Over a number of iterations, agents learn to adopt one of two conventions: yielding to the left, or yielding to the right. The authors quote Axelrod on the self-enforcing nature of norms (see section 1.1 above). In fact the "rules of the road" scenario doesn't fit his definition all that well. The model does not include punishment of those who are "seen" to drive on the wrong side of the road, rather the negative sanctions only arise when a driver collides with an oncoming vehicle or stops because his way is blocked; and these consequences are of course equally costly for the "conformist" and the "deviant". In fact this study provides a nice example of the difference between a convention and a norm: there is no element of normative appraisal, i.e. there is nothing in Sen and Airiau's setup which empowers agents to administer sanctions for observed infringements, and there is no discussion of how such a normative framework could emerge once a convention is in place. Elster [6] points out that "rules of the road" may function simultaneously as conventions and legal norms – drivers may tacitly agree to keep to one side of the road to avoid crashing into each other, even if they are not on a public highway or there are no traffic police within 100 miles.

3.3 SAVARIMUTHU, CRANEFIELD, PURVIS AND PURVIS (2010). OBLIGATION NORM IDENTIFICATION IN AGENT SOCIETIES.

Savarimuthu et al. [20] present a model which is intended to simulate an agent's acquisition of norms in an unfamiliar environment. This model involves two main functions: norm identification and norm verification. The scenario is that the agent (let's call him the **diner**) is visiting a restaurant in a strange country, and is naturally anxious to know how people are expected to behave when eating out in this country; specifically, whether or not he should leave a tip for the waiter. (Tipping etiquette is a popular topic in studies of normativity, no doubt because scholars are regularly confronted with this problem on conference trips.) The diner is supposed to observe a series of episodes involving tippers and non-tippers, and apply data mining techniques to discover if sanctioning actions are reliably associated with the presence of absence of any identifiable sequence of events. This leads to the postulation of candidate norms which are verified by questioning a local agent. Under certain assumptions the system does indeed succeed in learning that tipping is expected. There are (at least) two considerations here: firstly, for tipping to count as a norm, the waiters' actions should also be considered appropriate within the society – there should be a permissive norm for waiters to react angrily to non-tipping customers, and this is something that may be done correctly or incorrectly. And secondly, the diner needs to correctly interpret the waiters' actions as sanctions. However, in this model sanctioning actions are considered to be transparent, and the waiters perform them "probabilistically" rather than under any kind of accountability. Also: a customer's decision not to tip may itself count as a "sanctioning action" if the customer is not satisfied with their service. However, the diner cannot ascertain this unless he already knows whether a tipping norm is in place – if it is not, then failure to tip carries no significance as a sanction.

We have already mentioned the "gerrymandering" problem above: in order to correctly identify a norm of tipping, an agent must assume that there is no other plausible explanation for a waiter's chastising a customer. Yet as has been argued, any sequence of events may in principle be interpreted as instantiating a variety of regularities: there are for example many norms governing proper behaviour in restaurants, such as dressing appropriately, not getting drunk and raucous, not using one's mobile phone and so on, and a breach of any of these norms could lead to a customer being chastised by a waiter. However, the agents in this model are in fact equipped with an extremely parsimonious ontology: the only customer actions which can be perceived are {*arrive*, *order*, *pay*, *tip*, *depart*} while waiters have just two actions, *wait* and *sanction*.

In other words, an outside observer can't simply try to infer norms by looking out for sanctioning actions, as the local norms themselves determine what counts as a sanction and whether it is properly applied. A second conclusion is that norms are manifested in interactions that exhibit mutual accountability: if either party decides to sanction the other, this only makes sense if (a) the sanctionee both understands the significance of the action and accepts it as appropriate (b) the sanctioner acts deliberately, and is prepared to explain and justify his action. There seems to be some partial recognition of this issue in the same authors' [19] which suggests that "punishing agents can communicate the reason for punishment to the agents that ask for norm verification", though this does not in fact seem to be implemented in the model and there is no mechanism for challenging or appealing against punishments. The authors concede that "recognising and categorising a sanctioning event is a difficult problem" but assume "that such a mechanism exists (e.g. based on an agent's past experience)". Given that sanctioning is itself a norm-governed activity, it seems (as argued by [14]) that the authors are assuming that what they are seeking to explain is already understood: the "diner" has already somehow acquired an understanding of sanctioning norms.

3.4 CAMPENNI, ANDRIGHETTO, CECCONI AND CONTE, 2009: NORMAL = NORMATIVE? THE ROLE OF INTELLIGENT AGENTS IN NORM INNOVATION.

This paper explicitly differentiates norms from "mere conventions" and models the spread of normative beliefs within a population using the normative architecture EMIL-A. The authors acknowledge that this on its own is not sufficient to explain norm-conformant behaviour, as agents need to not only recognise normative beliefs but adopt them and act as if bound by them, but they follow a divide-and-conquer strategy of deferring this problem for another occasion. Essentially, agents exchange messages which encode various types of normative belief, such as "It is polite to answer when asked"; these may be expressed using deontic commands, evaluative statements, assertions about the state of the world or requests. Unlike many other analyses, this approach recognises that normative beliefs may be understood and acquired through many different channels, from explicit commands to observation of exemplary behaviour. A threshold determines how frequently an agent needs to observe a particular normative behaviour before adopting the corresponding normative belief. The reader may already have noted that the above example assumes a pre-existing notion of "politeness", itself a normative concept. And although this approach is contrasted with studies based on "behavioural regularities" it still ultimately rests on quantitative considerations by using a numerical threshold. It is not clear what would happen if the threshold is less than 50% and two conflicting norms are both observed with the same frequency. The reported simulation uses a threshold of 99% which is surely unrealistic for many real-word scenarios. Also, reliance on a threshold is rather an austere idealisation from actual social practice as people are also likely to be influenced by argumentation and by the perceived status and/or reliability of their informants.

3.5 BOELLA, COLOMBO TOSATTO, D'AVILA GARCEZ, GENOVESE, PEROTTI AND VAN DER TORRE, 2012. LEARNING AND REASONING ABOUT NORMS USING NEURAL-SYMBOLIC SYSTEMS.

This paper seeks to integrate symbolic and quantitative methods by training a neural network to learn rules encoded in a formalism called Input/Output Logic, which is essentially a list of condition-action rules. The system is provided with a subset of rules from the Robocup tournament and is then trained with instances of match behaviours, including actions which have been punished by the referee, with the objective of learning the remaining rules. Perhaps unsurprisingly, the performance increases directly according to the number of rules pre-encoded in the KB. Without discussing the results in detail, this simulation like that of [20] relies on a normatively-constituted notion of punishment and so may be prone to regress issues. This model cannot straightforwardly explain how the referee comes to know the rules, and appears to assume the referee acts correctly in all circumstances. The authors have an idiosyncratic use of terminology, classing the tournament rules as "regulative" as contrasted with "constitutive": according to Searle's original distinction [21], rules for a competitive sport are constitutive as they essentially create or define new forms of behaviour rather than regulating existing activities. In fact the examples discussed in the paper seem to combine constitutive rules of this kind with prescriptive rules such as *IF have_ball and opponent_approaching THEN O(pass)*.

3.6 SUMMARY

To summarise this section, we have considered a selection of agent-based models which aim to model the emergence of norms through various forms of learning, either through observation of agents who are assumed to exemplify the behaviour in question or as a result of positive or negative reinforcement. I claim that the rules-of-the-road example, [22], does not have to do with norms as such but with Lewisian conventions, as there is no real element of normative appraisal or accountability. I would argue that each of the remaining approaches is either liable to regress problems or manifests some degree of "gerrymandering". Both [2] and [20] involve the learning agent observing episodes of sanctioning behaviour, which as previously noted implies both that the agent has a prior notion of what constitutes "sanctioning" within this community, and that it can tell whether or not the sanctions are correctly applied. As noted, [5] assumes that the agent has a prior notion of "politeness". And one does not need to subscribe to an extreme scepticism to acknowledge that these models tend to operate with rather simplistic models of individuals and patterns of behaviour.

4. DISCUSSION, RELATED WORK AND FUTURE DIRECTIONS

In this suggestion we briefly suggest some implications for the design of agent-based simulation models arising from the above considerations, and consider possible directions for future research.

4.1 COMMUNICATIVE RATIONALITY

If we interpret normativity in terms of mutual accountability, following e.g. [3, 4, 9, 18], then agent-based modelling will require more than probabilistic reasoning, machine learning and signalling between agents; agents need to have "communicative competence" in the sense of being able to challenge or justify any sanctioning actions or responses to sanctions. This aspect seems to be missing from the "normative process model" proposed by [12]. Their model of the norm life cycle includes: creation, transmission, recognition, enforcement, acceptance, modification, internalisation, emergence, forgetting and evolution. However, there seems to be no recognition of the part played in these processes by negotiation and argumentation, which would seem essential, for example, for assessing whether sanctions are appropriately applied and challenging misapplications. (A survey of the state-of-the art in argumentative agents can be found in [24]) A reviewer noted that this raises the question of whether ABM systems would "need to be endowed with the intelligence, culture and linguistic abilities of humans" or whether a "lower level of competence" would suffice. This is a question to which I have no immediate answer but which could stimulate a fruitful research programme.

4.2 REGRESS PROBLEMS

We have noted that attempts to simulate the emergence of norms in multi-agent systems are prone to various types of regress problem, such as:

1. Any attempt to explain emergent properties of an agent community itself relies on previous norms and rules – it is simply not possible to start with a "state of nature" that is free of any assumptions or preconceptions.
2. Norms come as part of a package that includes "metanorms": if an agent applies sanctions to enforce a norm, or fails to sanction deviance from a norm, that action or inaction is itself potentially subject to normative appraisal, and so on recursively.

In response to point (1): developers of simulation models need to recognise that the initial state of the model will unavoidably encode implicit or explicit assumptions stemming from design decisions, and it is essential to be clear about what these assumptions are and to justify them. Agents' preferences are determined by institutionalized values as much as by their own "desires".

Brandom [3] discusses point (2) but does not offer a clear solution. Rouse's approach [18] is a little mysterious. He aims to show that "[a] normative conception of practices makes normativity irreducible but not inexplicable" and locates normativity in patterns of behaviour which are characterised by their mode of interaction rather than any observable regularities or prior meanings. The key feature is that "these patterns of interaction constitute something at issue and at stake in their outcome". I must admit to finding his line of argument somewhat opaque and in need of further investigation.

Heath [9: 114] suggests what he calls "a promising strategy for solving the regress problem" which is to treat mutual accountability as a symmetric relation between just two agents: one person acts, the second approves his action (or not), the first person assesses the propriety of the second agent's response and so on. It has to be said that this seems a little unrealistic, and assumes a benign cooperativity on the part of both players: in day-to-day encounters if one chastises someone for "deviant" behaviour such as parking on the pavement or using the wrong check-out till at a supermarket, their response may well be testy and angry rather than a reasoned appraisal of the rights and wrongs of the matter. In such cases it is likely that a third party would be called in to resolve the issue.

While the regress argument seems to rule out various classes of explanations for the origin of norms, it does not seem to leave space for any satisfying alternative. The idea that we should give up on the idea of terminating the regress and accept a kind of "norm holism" feels uncomfortably close to admitting defeat. Turner [25] attacks this issue by arguing that the "normativists" (his term) like Brandom, Sellars and Rouse are going about things the wrong way and proffering solutions for non-problems. Normativist analyses are actually going astray by trying to apply the methods of philosophy in what is properly the domain of social theory. According to Turner, any theory that purports to account for normative "facts" such as obligations, correctness, validity and so on is doomed to collapse into circularity, which can only be terminated by postulating fictions such as "collective beliefs" or a *Grundnorm* (basic norm). He rejects the notion that there are any such "facts" to be accounted for and proposes that normative phenomena are best accounted for using the methods of social science, investigating the beliefs that people hold about what is proper, correct or appropriate, without the analyst needing to endorse any of these notions.

This still seems to leave us with the question of how our normative beliefs tend to be so closely aligned. It turns out that Turner has a regress-stopper of his own, which he claims to be more legitimate and scientifically respectable: *empathy*. We all share the same biological make-up, and are apparently endowed with "mirror neurons" which may predispose us to understand and imitate other people's intentional actions (op cit. 175-77, 204-5). This, it is claimed, is what enables us to converge on beliefs about what sort of behaviour is appropriate or legitimate in different social contexts. This is evidently an area where further research might well be illuminating, but it is not really possible to assess these claims on the basis of Turner's rather sketchy exposition.

An alternative approach to regress (2) might be to relax the problem rather than attempt a definite solution. Recall that [7] questioned Axelrod's decision to set meta-punishments at the same level of "vengefulness" as the punishment for the original offence: they proposed that the cost of meta-punishment for failing to sanction an offence should be one-tenth of the cost of the original punishment of the offence itself. Generalising this, a reasonable conjecture might be that both the likelihood of being sanctioned and the cost of being punished diminish at each level of appraisal, so that the impact becomes vanishingly small after a certain number of iterations and may be effectively disregarded. Elster [6] points out that "[s]ocial life simply does not have this relentless transitivity" and expresses severe doubts that third- or fourth-party observers would be inclined to initiate meta-punishment of non-punishers.

6. CONCLUSION

We have argued that certain representative studies of normative agency using agent-based simulations are limited in that they fail to account for the dimension of mutual accountability, which has been extensively discussed in the relevant philosophical literature. It has also been argued on the basis of regress problems that norm emergence can only be modelled against a background of existing norms and values, rather than assuming a pristine "state of nature". We have concluded with some tentative suggestions for incorporating these insights in the design of simulation models and outlined some directions for future research.

REFERENCES

[1] AXELROD, R (1986). An evolutionary approach to norms. *The American Political Science Review*, Vol. 80, No. 4 (Dec., 1986), pp. 1095-1111

[2] BOELLA, G, Colombo Tosatto, S, d'Avila Garcez, A, Genovese, V, Perotti, A and van der Torre, L, 2012. Learning and Reasoning about Norms using Neural-Symbolic Systems. Proceedings of AAMAS 2012, pp. 1023- 1030.

[3] BRANDOM, R (1994) *Making It Explicit: Reasoning, Representing, and Discursive Commitment.* Harvard University Press, Cambridge, MA.

[4] BRANDOM, R (2000) *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, Cambridge, MA.

[5] CAMPENNI, M, Andrighetto, G, Cecconi, F and Conte, R, 2009: Normal = Normative? The role of intelligent agents in norm innovation. *Mind Soc* (2009) 8:153-172.

[6] ELSTER, J (2009). Norms. In Hedström, P and Bearman, P (eds), *The Oxford Handbook of Analytical Sociology*, Oxford University Press, pp. 195 – 217.

[7] GALAN, M., and Izquierdo, L. (2005) Appearances can be deceiving: Lessons learned Re-Implementing Axelrod's 'Evolutionary Approach to Norms. *Journal of Artificial Societies and Social Simulation* 8 (3) 2 http://jasss.soc.surrey.ac.uk/8/3/2.html.

[8] HABERMAS, J (1981) *Theory of Communicative Action*.

[9] HEATH, J (2003) *Communicative Action and Rational Choice.* MIT Press.

[10] HEATH, J (2008) *Following the Rules: Practical Reasoning and Deontic Constraint.* Oxford University Press.

[11] HODGSON, GM (2007), Meanings of Methodological Individualism. *Journal of Economic Methodology,* **14**(2), June, pp. 211-26.

[12] HOLLANDER, CD and Wu, AS (2011) The current state of normative agent-based systems. *Journal of Artificial Societies and Social Simulation*, 14, 2011. http://jasss.soc.surrey.ac.uk/14/2/6.html.

[13] KIBBLE, R (2006) Speech acts, commitment and multi-agent communication. *Computational and Mathematical Organization Theory* 12: 127–145.

[14] KIBBLE, R (2012). Conformist imitation, normative agents and Brandom's commitment model. Procs of AISB 2012 symposium, *The Social Turn*, University of Birmingham July 2012.

[15] LEWIS, D (1969) *Convention: A Philosophical Study.* Harvard University Press.

[16] MIKES, G (1949) *How to be an Alien*. Andre Deutsch.

[17] NEUMANN, M (2008) Homo Socionicus: a Case Study of Simulation Models of Norms *Journal of Artificial Societies and Social Simulation* vol. 11, no. 4 6 http://jasss.soc.surrey.ac.uk/11/4/6.html

[18] ROUSE, J (2007) Social practices and normativity. Division I Faculty Publications. Paper 44. Wesleyan University.

[19] SAVARIMUTHU, BTR, Cranefield, S, Purvis, MA and Purvis, MK, (2010a) Norm identification in agent societies. *The Information Science Discussion Paper Series*, number 2010/03. University of Otago.

[20] SAVARIMUTHU, BTR, Cranefield, S, Purvis, MA and Purvis, MK, (2010b) Obligation norm identification in agent societies. *Journal of Artificial Societies and Social Simulation*, 13. http://jasss.soc.surrey.ac.uk/13/4/3.html.

[21] SEARLE, J (1971) What is a Speech Act? In Searle (ed), *The Philosophy of Language*, Oxford Readings in Philosophy, OUP, pp. 39-53.

[22] SEN, S and Airiau, S (2007) Emergence of norms through social learning. In *Procs of IJCAI07*, pages 1507 – 1512, 2007.

[23] SINGH, M, 1999 A social semantics for agent communication languages. *Proceedings of IJCAI workshop on agent communication languages (IJCAI-99)*, pp. 75 – 88.

[24] TONI, F (2010) Argumentative agents. In *Procs of IMCSIT*, pages 223–229. IEEE, 2010.

[25] TURNER, S (2010) *Explaining the Normative*. Polity Press.

[26] WINCH, P (1958) *The Idea of a Social Science and its Relation to Philosophy*. Routledge and Kegan Paul, London.

MoK: Stigmergy Meets Chemistry to Exploit Social Actions for Coordination Purposes

Stefano Mariani[1] and **Andrea Omicini**[2]

Abstract. Socio-technical systems are becoming increasingly complex mostly due to the unpredictability of human interactions. Furthermore, they typically work within Knowledge Intensive Environments (KIE), hence they need to deal with huge amounts of data. Coordination models are meant to cope with the increasing complexity of software systems, mostly due to the unwanted non-determinism generated by the interaction within complex systems. In this paper we describe how social actions – performed by agents interacting in a shared environment – can be exploited by a novel model for the coordination of KIE, by adopting both a nature-inspired and a cognitive/behavioural standpoint.

1 Introduction

Socio-technical systems – that is, systems in which human interaction plays a central role – are becoming increasingly complex and thus difficult to design, mostly due to the unpredictability of human behaviour [14]. Furthermore, such systems are often *Knowledge Intensive Environments* (KIE) [1], that is, they are meant to store, compute, and make a huge amount of (possibly heterogeneous) information accessible, together with its links to other information—either belonging to the same system or not [19].

Probably the most enlightening example of this kind of systems is the so-called Web 2.0 [18]: there, a massive amount of raw data, structured information, and organised knowledge is continuously produced, consumed, and shared by single individuals, social communities, and business companies, each one with a different aim. Other examples of mashup of heterogenous data, (inter-)actions, agents, and goals are social networks, such as FaceBook: again, different agents (individuals, interest groups, companies) share different kinds of information (posts, pictures, videos, hyperlinks) with the purpose of achieving a different effect (inform, advertise, learn).

Coordination models, languages, and infrastructures are usually adopted to cope with the increasing complexity of software systems [4, 20, 16] – parallel first, then concurrent, finally distributed –, mostly due to the growth of the *interaction space* that such systems have to manage *internally* – between the entities composing the system – and *externally* – toward their *environment* –, including other software system and humans. Such interactions have been for long recognised as an undesirable source of (uncontrollable) *non-determinism*, harming correctness, reliability, and predictability of systems.

Only recently a new trend emerged in the coordination community, pushing toward a re-interpretation of such non-determinism in an attempt to make the uncontrollable, well, controllable. Then, *probability* and *stochasticity* entered the picture as a means to model, govern, and predict non-determinism – making it become a source of solutions rather than problems –, often thanks to the adoption of *nature-inspired metaphors* [15]. In fact, real-world natural systems – traditionally belonging to chemistry, biology, physics, sociology, and the like – are widely recognised for their capability to "reach order out of chaos" through *adaptiveness* and *self-organisation*. These are exactly the features that novel coordination models bring into complex software systems – such as socio-technical and knowledge-intensive ones –, so as to both decrease the uncertainty of humans/agents interaction, and ease the management of huge amounts of data/processes.

To this purpose, we learned from natural systems that a few "primitive" capabilities have to be provided by the (coordination) system at hand:

Probability — Probability theory is an effective means to deal with non-determinism – both in a *descriptive* and a *prescriptive* way –, thus, in the end, to better understand and design the interaction space.

Time — Many natural systems have the ability to both recognise and react to the passage of time, which may impact their behaviour, and thus have to be accounted for.

Space — Lastly, almost every natural system features the ability to both recognise and react to changes in the spatial context it is living in, thus properly adapting its configuration and/or behaviour so as to better deal with the new environment.

Furthermore, the combination of time and probability lead to the *stochasticity* we observe in nature, and the combination of time and space produces the *awarness* exhibited by many natural systems.

In the remainder of this paper we first introduce a biological, cognitive, and behavioural (social) interpretation of coordination, and describe a (bio)chemically-inspired coordination model (Section 2); then we discuss how a novel model for the coordination of information in KIE could exploit social actions performed by users to (biochemically) self-organise information (Section 3); finally we share our computational vision about the future of socio-technical and knowledge-oriented systems (Section 4).

2 Stigmergy in Natural and Artificial Systems

In order to properly engineer an artificial system mimicking a natural one, we should firstly describe the natural metaphors adopted to conceive and design our self-organising coordination model (Subsection 2.1); then, the computational requirements needed to support such metaphors have to be recognised and described (Subsection 2.2), to be finally exploited (Subsection 2.3).

[1] Università di Bologna, Italy - email: s.mariani@unibo.it

[2] Università di Bologna, Italy - email: andrea.omicini@unibo.it

2.1 (Cognitive) Stigmergy & Behavioural Implicit Communication

Stigmergy

One of the fundamental factors driving the (self-)organisation of complex social systems – such as human organisations, animal societies, and multi-agent systems – is that interactions between individuals is *mediated by the environment*, which "records" all the *traces* left by agents actions [25]. Trace-based communication is related to the notion of *stigmergy*, firstly introduced in the biological study of social insects [6], e.g. to characterise how termites (unintentionally) coordinate themselves during the construction of their nest, with no need of exchanging direct messages, instead relying solely on *local interactions* [21]. There, the trace to follow by each termite is the evolving shape of the nest (perceived in a small neighbourhood), implicitly suggesting where to put a new brick and when to merge two different buildings.

Cognitive Stigmergy

A number of relevant works in the field of cognitive sciences point out the role of stigmergy as a fundamental coordination mechanism, especially in the context of human societies and organisations. There:

- modifications to the environment are often amenable of a *symbolic interpretation*, in the context of a shared, conventional system of *signs*;
- interacting agents feature *cognitive abilities* that can be proficiently exploited in the stigmergy-based coordination.

When traces becomes signs, stigmergy becomes *cognitive stigmergy*. There, self-organisation is based on signs, which require symbolic interpretation capabilities, hence involves *intelligent agents*, able to correctly understand traces as signs intentionally left in the environment—and to react properly [14].

Behavioural Implicit Communication

Another step beyond cognitive stigmergy – and thus stigmergy –, is the argument that self-organising coordination among agents can be based on the observation and interpretation of actions as wholes, rather than solely of their effects on the environment—be them either traces or signals. This is what is called *Behavioral Implicit Communication* (BIC), where communication does not occur through any specialised signal, but through the practical behaviour observed by the recipient [2]. Then, actions themselves – along with their traces, as usual – become the "message", often intentionally sent through the environment in order to obtain collaboration, either by the environment itself or by other agents.

2.2 Computational requirements

Moving from stigmergy to BIC, a list of *desiderata* emerge which a coordination system should satisfy in order to properly model the natural mechanisms for self-organising coordination.

Stigmergy — For plain stigmergy, a number of features should be supported, both regarding environment reactions to agent actions, and its structure:

- the "recording" of agent traces should be possible;
- proper reaction to the emission of such traces—e.g., pheromone-like traces should interact with the environment where they are deposited so as to evaporate;
- traces should be available to other agents for perception;
- furthermore, the environment should feature a *topology*, that is, a coherent and expressive set of spatial abstractions to both describe and manage *locality* of actions.

Cognitive Stigmergy — In addition to the above properties:

- cognitive stigmergy welcomes tools supporting agents in the symbolic interpretation of traces as signs, such as dictionaries and *ontologies*;
- agent *intelligence* is a necessary pre-condition—in contrast to stigmergy.

BIC — Being a generalised form of stigmergy-based communication, other features are needed to support BIC:

- agent actions aimed at behaviourally expressing coordination issues need to be made *observable* to other agents sharing the same environment;
- relevant properties of such actions need to be visible, too—e.g. when an action was performed, where and who did it.

Tuple-based Coordination

Among the many sorts of computational models for coordination [4], tuple-based ones [3] can be taken as a reference for stigmergic coordination—including cognitive and BIC [11]. There, multiple tuple spaces physically/logically distributed in a computational system could be seen as the building blocks of the system environment—supporting a first notion of topology related to network connections. Tuples are the information chunks stored by agents/process into tuple spaces, which in principle could reify any kind of (possibly, heterogeneous) data—thus, our traces too. Coordination primitives could then be used either by "stupid" processes to synchonise upon predetermined tuple patterns (templates), or by intelligent agents properly interpreting the symbolic content of tuples—as in the case of *logic tuples*, for instance [11].

Furthermore, extended tuple-based models (and infrastructures), such as TuCSoN [17], feature space-time awareness (hence leveraging topology too), probability and "event-driven" programmability – thanks to the ReSpecT language [12] –, providing us with all the necessary tools to fully support stigmergy, cognitive stigmergy, and BIC. Not by chance, TuCSoN is the coordination model and infrastructure chosen for a prototype implementation of the model discussed in Section 4.

2.3 (Bio)Chemistry & Biochemical Tuple Spaces

Among the many diverse natural metaphors available – physical, biological, social, etc. –, the chemical one appears particularly interesting for the simplicity of its foundation. The basic idea is to coordinate components (data and processes) as molecules floating in a solution (the distributed system), with *chemical rules* consuming and producing such molecules to drive the (self-organising) coordination process. As many chemical reactions can occur at a given time, system evolution is driven by their chemical *rate*, *probabilistically* selecting certain behavioral paths over others [5].

Biochemical tuple spaces are a stochastic extension of the LINDA framework [3]. The idea is to attach to each tuple a "concentration", which can be seen as a measure of the *pertinency/activity* of the tuple—the higher it is, the more likely and frequently the tuple will influence system coordination [24]. Concentration of tuples is dynamic, as it evolves thanks to chemical rules that can be installed into the tuple space, which affect concentrations over time precisely in the same way chemical substances evolve into chemical solutions. Interaction between tuple spaces is achieved through a special kind of chemical law, which "fires" some tuples to a tuple space in the *neighbourhood*, picked probabilistically. This mechanism mimics the concept of biological compartment, whose boundary can be crossed by chemical substances, thus allows to conceive systems as networks of nodes.

Ultimately, biochemical tuple spaces embeds the most advanced features exhibited by natural systems—thus also the more primitive ones highlighted in the Introduction:

Stochasticty — Since chemical laws are meant to be executed as an exact simulation of chemical solutions' dynamics, every single reaction is executed probabilistically based on its rate, also influencing the time needed to complete execution.

Space-Time Awarness — The neighborhood notion and its bond with probability aspects, completes the picture by leveraging the chemical metaphor – considering only one solution in isolation – to a biochemical one, allowing description and management of localities and contingencies typical of natural systems.

In fact, e.g., by interpreting each tuple as a service – actually, as the reification of the relevant properties of a service— and by installing adequate chemical laws, biochemical tuple spaces support a significant extent of the following self-organisation properties:

Self-adaptation — only the best services survive among a set of competing ones, such a "natural selection" depending on the dynamics of agents' incoming requests, rather than on external, deterministic control;

Spatial-sensitiveness — the firing tuples mechanism makes competition be a spatial notion, possibly partitioning the network into *ecological niches* where different services develop better than others;

Openness — the same set of laws are expected to work in spite of the unpredictable incoming of new types of services and request.

All these desirable features provided by biochemical tuple spaces, make such model a suitable ground upon which to build our own self-organising knowledge environment model, presented in next Section.

3 The Molecules of Knowledge Model

The Molecules of Knowledge model (MoK for short) was introduced in [9] as a novel framework to conceive, design, and describe knowledge-oriented, self-organising coordination systems. An early application of its principles has also been showed in [8], taking news management as a reference case study.

Briefly, the main ideas behind the MoK model are that:

- existing information should *autonomously link* together, progressively clustering into more complex heaps of knowledge;
- rather than searching for interesting information, users of the system – either humans or agents – should see such information *spontaneously manifest* to and *diffuse* toward them.

In order to do so, (bio)chemistry was taken as a source of inspiration, thus biochemical tuple spaces as a reference model to specialise. Then, after a short overview of the model clarifying its biochemical roots (Subsections 3.1–3.3), we provide a novel, BIC-oriented interpretation of the same (Subsection 3.4).

3.1 MoK Overview

MoK main abstractions are:

Atoms — the smallest unit of information, it contains information from a knowledge source and belongs to a compartment where it "floats";

Molecules — MoK heaps for information aggregation, they cluster together somehow related atoms;

Enzymes — emitted by catalysts (see below), enzymes represent the *reification* of knowledge-oriented, possibly epistemic social (inter-)actions, and are meant to influence the dynamics of information evolution by participating in MoK reactions;

Reactions — working at a given rate and applied to given atoms/molecules, reactions are the biochemical laws regulating the evolution of the shared environment, by governing information aggregation, diffusion, and decay within MoK compartments (below).

Furthermore, aspects like topology, knowledge production and consumption are addressed by the following further abstractions:

Compartments — the spatial abstraction of MoK, they represent the conceptual *loci* for all other MoK abstractions – atoms, molecules, etc. –, thus provide MoK with the notions of locality and neighbourhood;

Sources — each one associated to a compartment, MoK sources are the origins of knowledge, which is injected in the form of atoms within the compartment they belong to;

Catalysts — the abstraction for knowledge prosumers (producers + consumers), catalysts (unintentionally) emit enzymes whenever they interact with/throug their compartment, in order to reach their own goals.

The formal definition of MoK main entities follows in next Section—syntax is slightly modified w.r.t. [9] to ease understanding.

3.2 Formal MoK

Atoms

Being produced by a knowledge source and conveying a primitive piece of information, atoms should also store some contextual information to refer to the content origin and to preserve its original meaning. As a result, a MoK atom is essentially a triple of the form:

$$\texttt{atom}(src, val, attr)^{c}$$

where val is the information chunk to be stored whereas $attr$ is any kind of *ontological* meta-information useful for both the MoK system and MoK users to better handle the atom. Superscript c is the atom concentration value, inherited from the biochemical tuple space model.

Molecules

A MoK system can be seen as a collection of atoms wandering in and between compartments, possibly colliding with each other. The result of collisions are the molecules of knowledge, that is, spontaneous, stochastic, "environment-driven" aggregations of atoms, which are meant in principle to reify some semantic relationship between atoms, thus possibly adding new knowledge to the system—e.g., stating that this information is related to that. Hence, each molecule is simply a set of atoms:[3]

$$\texttt{molecule(}\mathit{Atoms}\texttt{)}^{c}$$

Enzymes

One of the key features of MoK is that the system interprets prosumer's (*epistemic*) actions as positive feedbacks, increasing the concentration of involved information chunks within the prosumer's workspace. In order to do so, whenever catalysts somehow access molecules within their compartment, a number of enzymes is released to reify such action. Then, MoK biochemical reactions consume such enzymes to properly increase molecules' concentration, ultimately enforcing the positive feedback typical of natural, self-adaptive systems.

Being bound to the catalyst compartment, the minimal information an enzyme has to veichle is the accessed information. The structure of an enzyme is then:

$$\texttt{enzyme(}\mathit{Molecule}\texttt{)}^{c}$$

Biochemical Reactions

The behaviour of a MoK system is actually determined by biochemical reactions, which drive atoms and molecules aggregations, as well as reinforcement, decay, and diffusion.

As a knowledge-oriented model, the main issue of MoK is determining the *semantic correlation* between information. So, to completely design a MoK system, one should first of all define the basic MoK *function*, taking two molecules and returning a value $m \in [0, 1]$ reflecting the degree of similarity between them:

$$\mathcal{F}_{mok}\colon \mathit{Molecule} \times \mathit{Molecule} \longmapsto [0, 1].$$

Then, MoK biochemical reactions can be defined by relying on the application-specific $\mathcal{F}_{mok}$:

Aggregation — The aggregation reaction bounds together atoms and molecules, based on their semantic correlation:

$$\texttt{molecule}(\mathit{Atoms}_1) + \texttt{molecule}(\mathit{Atoms}_2) \longmapsto^{r_{agg}}$$
$$\texttt{molecule}(\mathit{Atoms}_1 \biguplus \mathit{Atoms}_2) + \mathit{Residual}(\mathit{Atoms}_1 \bigcup \mathit{Atoms}_2)$$

where:

- superscript r_{agg} is the reaction rate;
- $\mathcal{F}_{mok} \geq 0 + \delta$ only for atoms $atom_i \in Atoms_1$, $atom_j \in Atoms_2$ with $i \neq j$, linked by the produced molecule—δ is an arbitrary threshold;
- and $Residual(Atoms_1 \bigcup Atoms_2)$ represents the set of $atom_i \in Atoms_1$, $atom_j \in Atoms_2$ for which $\mathcal{F}_{mok} < 0{+}\delta$, released in the compartment—so that the total number of atoms is preserved.

The outcome of any single application of the aggregation reaction, is the birth of a molecule meant to represent an existing semantic link between its consituent atoms.

Reinforcement — Positive feedback is due to the *reinforcement reaction*, which consumes a single unit of enzyme injected by a catalyst to produce a single unit of the targeted atom/molecule.

$$\texttt{enzyme}(\mathit{Molecule}_1) + \mathit{Molecule}_1^c \longmapsto^{r_{reinf}} \mathit{Molecule}_1^{c+1}$$

Being enzymes released contextually to a (epistemic) knowledge-oriented action performed by a certain MoK user, they reify – thus makes observable and traceable – the interest of that user about that information. Furthermore, being catalysts associated to their compartments, also the conceptual place where the action took place becomes parts of the action reification. The MoK model exploits then such implicit information to infer that the accessed knowledge must be reinforced within the catalyst compartment.

Decay — In order to provide the *negative feedback* required to close the feedback loop typical of natural systems, molecules should fade as time passes, lowering their own concentration according to some well-defined *decay law*. The temporal *decay reaction* is hence defined as follows:

$$\mathit{Molecule}^c \longmapsto^{r_{decay}} \mathit{Molecule}^{c-1}$$

Diffusion — Analogously, a distributed self-organisation model should provide some kind of spatial evolution pattern. According to its natural inspiration, MoK adopts *diffusion* as its knowledge migration mechanism: atoms and molecules can only migrate between *neighbour* compartments, resembling membrane crossing among cells. MoK *diffusion reaction* is then modelled as follows—assuming that σ identifies a biochemical compartment and $\|\|_\sigma$ to enclose molecules in σ:

$$\{\mathit{Molecules}_1 \bigcup \mathit{Molecules}_1\}_{\sigma_i} + \{\mathit{Molecules}_2\}_{\sigma_{ii}} \longmapsto^{r_{diffusion}}$$
$$\{\mathit{Molecules}_1\}_{\sigma_i} + \{\mathit{Molecules}_2 \bigcup \mathit{Molecules}_1\}_{\sigma_{ii}}$$

where σ_i and σ_{ii} are somehow defined to be neighbour compartments.

3.3 MoK Self-Organisation

By recalling the fundamental primitive features exhibited by natural systems and met by the biochemical tuple space abstraction, we can see how they are also met by the MoK model:

Time — In the spirit of biochemical tuple spaces, time-awareness is embedded within the model in two different ways: through the decay reaction – making molecules influenced by the passage of time – and thanks to the reaction scheduling process itself, being it rate-dependent;

Space — Spatial aspects are accounted for in a number of ways. First of all the assumption on a either physical or virtual topology defining neighbourhood relationships between compartments. Then, the diffusion reaction exploits such underlying notion to mimic biochemical diffusion across membranes—(logical/physical) links between compartments. Finally, the aggregation reaction embeds the notion of locality, since only molecules sharing a compartment can be aggregated together.

Probability — Other than in the stochasticity implicit in reaction execution – since only the reaction probabilistically-selected according to rate wins among competing ones –, probability is exploited at another level: the matching function μ. Such function is typical of LINDA models and is used to select which tuples (molecules, here) have to be selected when considering a tuple template (chemical reactants, here)—potentially matching more than one. In MoK, the matching function μ is $\mathcal{F}_{mok}$-based, that is a matching tuple is probabilistically selected among the possibly many by taking into account the value m returned by $\mathcal{F}_{mok}(t, T)$: the higher the m, the higher the likelihood to be retrieved.

[3] As done in [9], for the sake of simplicity, one could also see atoms as molecules with a single atom, hence whenever helpful we consider $a = m$ and talk generally about molecules, implicitly including atoms.

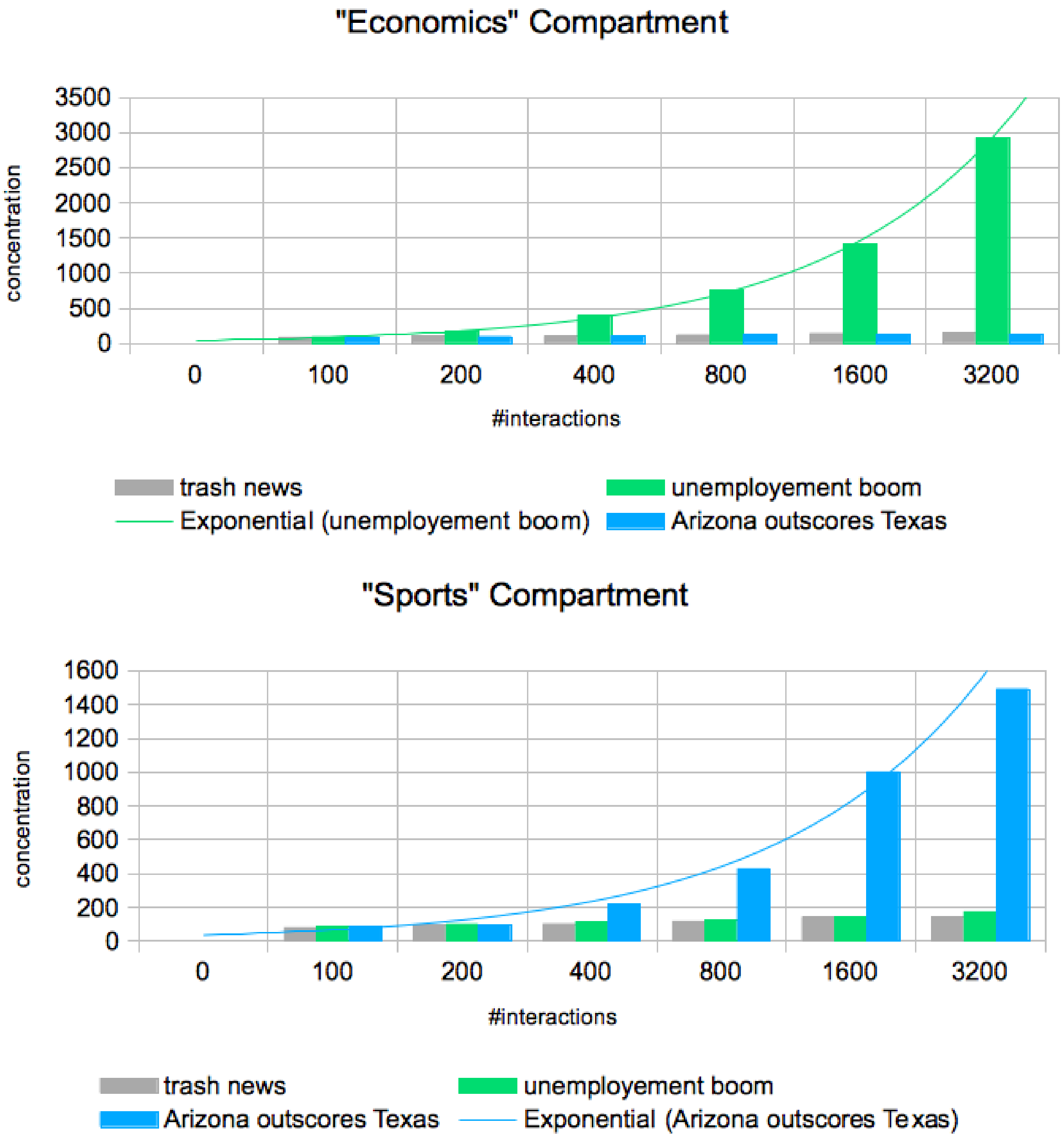


Figure 1. The stochastic equilibrium between diffusion, reinforcement and decay laws, makes a "smart diffusion" pattern appear by emergence.

These simple mechanisms altogether enable self-organisation of knowledge, in a similar way as that observed for services in biochemical tuple spaces. Next subsection shows a simple yet effective example of this, in the spirit of those provided in [8].

A Sample Application

The example consists of a MoK system dealing with the following scenario. A "producer" compartment stores a collection of different MoK sources – e.g. news articles talking about weather, baseball and finance – and simply diffuses them to the neighbour compartments "economics" and "sports"—belonging to journalists devoted to that particular topic. We expect that after a while the system would reach an "equilibrium" in which the two topic-oriented compartments are mainly populated by topic-compliant news molecules, whereas those not compliant should progressively tend to fade away. Furthermore, we also expect that the equilibrium is reached only as a consequence of journalists (catalysts) interactions—e.g. searching, collecting, exploiting, modifying information.

Figure 1 shows this actually happening.[4] As explained in [8]: dif-

[4] Actually, the simulation was run on an improved version of the prototype implementation of biochemical simulator used in [8]—althought still with the ReSpecT language [12] and the underlying TuCSoN coordination infrastructure [17].

fusion is implemented to be equiprobable towards each neighbour compartment; positive feedback is enacted by reinforcement reactions that take a molecule and the relative enzyme producing two – thus increasing concentration by one–; whereas negative feedback comes from the known decay law. While diffusion and decay rates are comparable, positive feedback was set higher.

The exponential growth observed is due to the influence that concentration of molecules has on the execution rate of biochemical laws: at equal rates, in fact, the law with higher concentrations of reactant molecules has actually a higher chance of being scheduled—in line with the (bio)chemical metaphor.

3.4 The MoK Model as a BIC Model

If we recall the computational requirements (highlighted in Section 2) needed to enable BIC-based coordination in artificial systems – thus (cognitive) stigmergy-based too –, we find the MoK model currently lacks three:

- making traces of agents' interactions available for observation to other agents;
- making agents' interactions themselves available for other agents inspection;
- explicitly record contextual information about such actions—e.g. who issued them.

In fact, the MoK mechanism closest to these features is the way in which enzymes are both produced and consumed. However, it is not enough for the following reasons:

- enzymes can represent easily represent both actions and their traces, but their observation is restricted to the environment, not made available to other agents—since they are consumed and cannot diffuse;
- again, contextual information is implicitly conveyed by enzymes – remember that each compartment belongs to a catalyst – but it is not available to other agents for inspection.

Nevertheless, a few simple extensions to the MoK model could be devised so as to better deal with BIC modelling. First of all, we can easily keep track of contextual information regarding agents interactions by simply associating each enzyme to a *descriptor* (σ) of the compartment they were released into[5]:

$$\texttt{enzyme}(\sigma, \mathit{Molecule})^{c}$$

Such a descriptor could store any meta-information about the compartment useful to better understand the action: its current time, place (think to a mobile device), the molecule with the highest concentration (e.g., to infer compartment topic of interest), and so on.

Then, we should make enzymes available for perception to other agents sharing the MoK system, and enhance them to better resemble traces of actions rather than actions as a whole. The simplest way to do so is to:

- allow them to participate in MoK diffusion reactions—so that both other agents and compartments can observe incoming enzymes, possibly deciding to do something with them;
- produce a "dead enzyme" whenever enzymes are involved in a reinforcement reaction, then allowing the dead to fade in time and diffuse in space.

The latter novelty actually makes it possible to *(i)* better distinguish traces of actions (dead enzymes) from actions (enzymes), *(ii)* make traces closer to what pheromones represent in natural systems, since they diffuse their scent depending on time and space patterns—actually intensity lowers as time passes and distance increase.

Reinforcement reaction should then be rewritten as:

$$\texttt{enzyme}(\sigma, \mathit{Molecule}_1) + \mathit{Molecule}_1^c \longmapsto^{r_{reinf}} \mathit{Molecule}_1^{c+1} + \texttt{dead}(\texttt{enzyme}(\sigma, \mathit{Molecule}_1))$$

whereas for the other extensions it is enough to let reactions decay and diffusion apply respectively to dead enzymes solely and both enzymes.

In next section, we aim at providing material for discussion, in the attempt to share our vision upon what next generation knowledge-oriented, socio-technical systems should look like, and how the social and cognitive theories of interactions could help coordination models and infrastructures cope with them.

4 Toward Self-Organising, Social Workspaces

The Molecules of Knowledge model can be thought of as a first prototype – actually implemented upon the TuCSoN middleware [22], and used for the tests in this paper – paving the way toward a much more complex and general idea of *self-organising workspaces* [13]. There, not only suitable methods and models have to be adopted to properly *engineer knowledge* – in particular, borrowing from knowledge representation and extraction techniques, e.g. conceived for the Semantic Web –, but also to support its self-organisation and adaptiveness w.r.t. an ever evolving working environment—as typically is for knowledge workers in general, e.g. researchers, journalists, lawyers, and the like.

In this context, two of the main concerns we could think of are already addressed by MoK – although in a rather primitive way –, that is, *knowledge aggregation* and *organisation*. The former is ensured by the molecule abstraction, actually reifying semantic relationships among different knowledge chunks. The latter is addressed by the combined contribution of diffusion, reinforcement, and decay, in which the enzyme abstraction plays a central role. Moreover, MoK extension toward BIC-based coordination could be useful to further push these mechanisms to their limits, hopefully realising forms of knowledge coordination driven by social actions and interactions.

In particular, a socio-technical system for knowledge-oriented coordination should properly handle *pervasive cognition* – that is, distributed cognition in pervasive computing scenarios –, where knowledge is pervasively distributed and is to be *accessible ubiquitously*—as witnessed by social services and networks like LinkedIn and Facebook. A foremost source of inspiration could come from the *data in the cloud* paradigm, which is based upon the idea of providing abstract interfaces for storing, retrieving and managing large amounts of data from anywhere, at any time, in the Web.

Accordingly, a novel *knowledge in the cloud* paradigm could be conceived, aiming at transforming data clouds into (technical and scientific) *semantic clouds*, spontaneously emerging, appearing, disappearing, splitting in different clouds, merging with one another and so on—as a consequence of the self-organising processes strongly relying on cognitive and semantic aspects also exploited in MoK. Any action of producing, accessing, consuming, elaborating, relating information would trigger local processes of elaboration that would result in the (self-)organisation of semantic clouds. This could be achieved by exploiting all the already-cited, well-known mechanisms ranging from stigmergy to Behavioural Implicit Communication.

[5] Note that a compartment descriptor has to be preferred over a catalyst because a catalyst can interact with more than one compartment – e.g. one for topic of interest – while a compartment strictly belongs to a catalyst.

5 Related Works

Other than already cited works on the Molecules of Knowledge model [9, 8], stigmergy and cognitive stigmergy [21, 14], and the BIC model [2, 23], a few more works from the coordination community are close to ours.

TOTA

The TOTA middleware was proposed in [7] as a self-organising, pervasive, tuple-based model, inspired by physics. There, each tuple is equipped by two additional fields other than its content:

- a *propagation rule*, determining how the tuple should propagate and distribute across the network of linked tuple spaces;
- a *maintenance rule*, dictating how the tuple should react to the passage of time and/or events occurring in the space.

The combination of the above rules enables the creation and self-adaptation of computational fields, that is, distributed data structures – such as *gradients* – enforcing spatio-temporal properties in the configuration of tuples, eventually exploited by coordinating agents. Through such fields, it is quite easy to implement most forms of stigmergy-based coordination.

SAPERE

SAPERE [26] is a biochemically-inspired model for the engineering of complex self-organising and adaptive pervasive service ecosystems. In SAPERE agents share LSAs (Live Semantic Annotation) – which could be thought of as a special kind of tuples – representing them and allowing them to interact in a shared environment while pursuing their own goals. LSAs are managed by the SAPERE ecosystem through eco-laws, which are biochemical-like rules responsible to evolve LSAs according to both agents and systems needs.

The complexity of both the LSAs syntax and semantics and that of eco-laws scheduling and execution policy makes SAPERE a very powerful model for general-purpose, self-organising, and adaptive coordination [10]. Furthermore, existence of full-fledged languages to both express and manipulate semantically LSAs as well as eco-laws, virtually enables any kind of BIC-based coordination pattern.

6 Conclusion

In this paper, we discuss a few well-known principles borrowed from cognitive and behavioural (social) sciences, and show how they can be exploited by computational systems – along with the biochemical metaphor – in order to better deal with knowledge-intensive environments.

Then, we describe the novel knowledge-oriented coordination model called Molecules of Knowledge, and discuss how it could be extended so as to better deal with BIC-related aspects.

Finally, we share some of our ideas regarding the future of socio-technical systems, hopefully paving the way towards new stimulating collaborations between the research fields of sociology, cognitive science, and coordination models and languages.

ACKNOWLEDGEMENTS

This work has been supported by the EU-FP7-FET Proactive project SAPERE Self-aware Pervasive Service Ecosystems, under contract no. 256873.

REFERENCES

[1] Ganesh D. Bhatt, 'Knowledge management in organizations: examining the interaction between technologies, techniques, and people', *Journal of Knowledge Management*, **5**(1), 68–75, (2001).

[2] Cristiano Castelfranchi, Giovanni Pezzullo, and Luca Tummolini, 'Behavioral implicit communication (BIC): Communicating with smart environments via our practical behavior and its traces', *International Journal of Ambient Computing and Intelligence*, **2**(1), 1–12, (January–March 2010).

[3] David Gelernter, 'Generative communication in Linda', *ACM Transactions on Programming Languages and Systems*, **7**(1), 80–112, (January 1985).

[4] David Gelernter and Nicholas Carriero, 'Coordination languages and their significance', *Communications of the ACM*, **35**(2), 97–107, (1992).

[5] Daniel T. Gillespie, 'Exact stochastic simulation of coupled chemical reactions', *The Journal of Physical Chemistry*, **81**(25), 2340–2361, (December 1977).

[6] Pierre-Paul Grassé, 'La reconstruction du nid et les coordinations interindividuelles chez Bellicositermes natalensis et Cubitermes sp. la théorie de la stigmergie: Essai d'interprétation du comportement des termites constructeurs', *Insectes Sociaux*, **6**(1), 41–80, (March 1959).

[7] Marco Mamei and Franco Zambonelli, 'Programming pervasive and mobile computing applications: The TOTA approach', *ACM Transactions on Software Engineering Methodologies*, **18**(4), 15:1–15:56, (July 2009).

[8] Stefano Mariani and Andrea Omicini, 'Self-organising news management: The *Molecules of Knowledge* approach', in *1st International Workshop on Adaptive Service Ecosystems: Natural and Socially Inspired Solutions (ASENSIS 2012)*, eds., José Luis Fernandez-Marquez, Sara Montagna, Andrea Omicini, and Franco Zambonelli, pp. 11–16, SASO 2012, Lyon, France, (10 September 2012). Pre-proceedings.

[9] Stefano Mariani and Andrea Omicini, 'Molecules of Knowledge: Self-organisation in knowledge-intensive environments', in *Intelligent Distributed Computing VI*, eds., Giancarlo Fortino, Costin Bădică, Michele Malgeri, and Rainer Unland, volume 446 of *Studies in Computational Intelligence*, pp. 17–22. Springer, (2013). 6th International Symposium on Intelligent Distributed Computing (IDC 2012), Calabria, Italy, 24-26 September 2012. Proceedings.

[10] Sara Montagna, Mirko Viroli, Matteo Risoldi, Danilo Pianini, and Giovanna Di Marzo Serugendo, 'Self-organising pervasive ecosystems: A crowd evacuation example', in *3rd International Workshop on Software Engineering for Resilient Systems*, volume 6968 of *Lecture Notes in Computer Science*, 115–129, Springer, Geneva, Switzerland, (29–30 September 2011).

[11] Andrea Omicini, 'On the semantics of tuple-based coordination models', in *1999 ACM Symposium on Applied Computing (SAC'99)*, pp. 175–182, New York, NY, USA, (28 February – 2 March 1999). ACM. Special Track on Coordination Models, Languages and Applications.

[12] Andrea Omicini, 'Formal ReSpecT in the A&A perspective', in *5th International Workshop on Foundations of Coordination Languages and Software Architectures (FOCLASA'06)*, eds., Carlos Canal and Mirko Viroli, pp. 93–115, CONCUR 2006, Bonn, Germany, (31 August 2006). University of Málaga, Spain. Proceedings.

[13] Andrea Omicini, 'Self-organising knowledge-intensive workspaces', in *Pervasive Adaptation. The Next Generation Pervasive Computing Research Agenda*, ed., Alois Ferscha, chapter VII: Human-Centric Adaptation, 71–72, Institute for Pervasive Computing, Johannes Kepler University Linz, Austria, (May 2011).

[14] Andrea Omicini, 'Agents writing on walls: Cognitive stigmergy and beyond', in *The Goals of Cognition. Essays in Honor of Cristiano Castelfranchi*, eds., Fabio Paglieri, Luca Tummolini, Rino Falcone, and Maria Miceli, volume 20 of *Tributes*, chapter 29, 543–556, College Publications, London, (December 2012).

[15] Andrea Omicini, 'Nature-inspired coordination models: Current status, future trends', *ISRN Software Engineering*, **2013**, (2013). Article ID 384903, Review Article.

[16] Andrea Omicini and Mirko Viroli, 'Coordination models and languages: From parallel computing to self-organisation', *The Knowledge Engineering Review*, **26**(1), 53–59, (March 2011). Special Issue 01 (25th Anniversary Issue).

[17] Andrea Omicini and Franco Zambonelli, 'Coordination for Internet application development', *Autonomous Agents and Multi-Agent Systems*,

2(3), 251–269, (September 1999). Special Issue: Coordination Mechanisms for Web Agents.

[18] Tim O'Reilly, 'What is Web 2.0: Design patterns and business models for the next generation of software', *Communications & Strategies*, **65**(1st Quarter), 17–37, (31 March 2007).

[19] Sascha Ossowski and Andrea Omicini, 'Coordination knowledge engineering', *The Knowledge Engineering Review*, **17**(4), 309–316, (December 2002). Special Issue: Coordination and Knowledge Engineering.

[20] George A. Papadopoulos and Farhad Arbab, 'Coordination models and languages', in *The Engineering of Large Systems*, ed., Marvin V. Zelkowitz, volume 46 of *Advances in Computers*, 329–400, Academic Press, (1998).

[21] H. Van Dyke Parunak, 'A survey of environments and mechanisms for human-human stigmergy', in *Environments for Multi-Agent Systems II*, eds., Danny Weyns, H. Van Dyke Parunak, and Fabien Michel, volume 3830 of *LNCS*, 163–186, Springer, (2006).

[22] TuCSoN. Home page. `http://tucson.apice.unibo.it/`, 2013.

[23] Luca Tummolini, Cristiano Castelfranchi, Alessandro Ricci, Mirko Viroli, and Andrea Omicini, '"Exhibitionists" and "voyeurs" do it better: A shared environment approach for flexible coordination with tacit messages', in *Environments for Multi-Agent Systems*, eds., Danny Weyns, H. Van Dyke Parunak, and Fabien Michel, volume 3374 of *LNAI*, 215–231, Springer, (February 2005). 1st International Workshop (E4MAS 2004), New York, NY, USA, 19 July 2004. Revised Selected Papers.

[24] Mirko Viroli and Matteo Casadei, 'Biochemical tuple spaces for self-organising coordination', in *Coordination Languages and Models*, eds., John Field and Vasco T. Vasconcelos, volume 5521 of *LNCS*, 143–162, Springer, Lisbon, Portugal, (June 2009). 11th International Conference (COORDINATION 2009), Lisbon, Portugal, June 2009. Proceedings.

[25] Danny Weyns, Andrea Omicini, and James J. Odell, 'Environment as a first-class abstraction in multi-agent systems', *Autonomous Agents and Multi-Agent Systems*, **14**(1), 5–30, (February 2007). Special Issue on Environments for Multi-agent Systems.

[26] Franco Zambonelli, Gabriella Castelli, Laura Ferrari, Marco Mamei, Alberto Rosi, Giovanna Di Marzo Serugendo, Matteo Risoldi, Akla-Esso Tchao, Simon Dobson, Graeme Stevenson, Yuan Ye, Elena Nardini, Andrea Omicini, Sara Montagna, Mirko Viroli, Alois Ferscha, Sascha Maschek, and Bernhard Wally, 'Self-aware pervasive service ecosystems', *Procedia Computer Science*, **7**, 197–199, (December 2011). Proceedings of the 2nd European Future Technologies Conference and Exhibition 2011 (FET 11).

Sociology and AI: Requirements and achievements for walking towards a cross-fertilization integration

Francisco J. Miguel Quesada[1]

Abstract. We attempt here an exercise of analytical deconstruction of the basic requirements of Sociology to develop scientific knowledge, produced as a tool to improve open discussion about the mutual contribution opportunities between participants in the SOCIAL.PATH symposium. In this preliminary version, a catalogue of the main explanatory foundations in Sociology is linked with a tentative list of AI actual developments and tools, in order to create a path to research integration among disciplinary communities.

1 INTRODUCTION

In his work "The Sciences of the Artificial", Herbert Simon wrote: *"A veridical picture of economic actors and institutions must incorporate the information processing limits set by their inner environments. The picture must also accommodate both the conscious rationality of economic decision makers and the unplanned but adaptive evolutionary processes that have molded economic institutions."* (1996: 49) [1].

In line with this, a long tradition in the social sciences claims that the explanation of any social phenomena should be "reduced" to individual decision-making (rational choice approach) or to meaningful action (sociological *subjetivism*), while another tradition claims for the scrutiny of the social ontology without dependence on behavioral sciences (social system theory and sociological *structuralism*). The debate between both approaches could be found at the very origins of Sociology, and continued all along the history of the discipline, generating and expanding one of the current major paradigmatic divisions within the sociological scientific community [2,3].

Simon's statement claims to establish a route-map, or a protocol, to guide the modeling of economic behavior as part of the wider class of social phenomena. This idea of an integrative effort towards interdisciplinary collaboration could be also applied to social scientists and AI & MAS experts, as the recent development in ABM computational sociology shows. But there are a number of challenges that has to be faced in the path to collaborating ventures. For instance, some terminological and conceptual differences -between disciplines- have to be adressed and clarify. Also, some problem-solving approach differences -between engineering and scientific practitioners- have to be integrated in a common research framework [4], as the problems raised due to a lack of a clear and shared definition of explanatory and methodological foundations in sociology.

To cope with the later of these challenges, the integrative effort could begin by generating, in an explicit and understandable format, a detailed catalogue of requirements to develop scientific social knowledge, side by side with state-of-the-art achievements and available tools developed in the AI & MAS domain.

From a personal perspective while working in Computational Sociology, there are two different strategies for engage in collaboration: (a) some sociological puzzle, or case study, is proposed to be solved by AI practitioners, or (b) some AI related methodology or tool is used by social researchers. In any case, the integration could be understood as a *"horizontal"* unity across disciplines, which promotes the problem solving for the specific case of study. But, the proposal of Simon could inspire another orientation, towards a *"vertical"* integration: apart from the positive *"horizontal"* effects, looking from each domain into the other could produce relevant improvements into each discipline. In this sense, R. Axelrod has claimed that while developing a formal modeling of a social phenomenon the social scientist has to pay more attention to fine-grain sociological explanations, and also that while helping to implement a computational simulation the social scientist could develop a deeper and more integrated understanding of the phenomena under scrutiny [5]. From the point of view of AI, many metaphors of the social and human behavior and operation, have been useful to increase the catalogue of tools and methodologies for AI & MAS developers, and to improve the integrate corpus of knowledge of the discipline.

	Main benefit
Horizontal	For the specific problem solution
Vertical	For each discipline development

Table 1. Effects of integration efforts across disciplines.

With this cross-fertilization effect in mind, that outperforms the simple *"horizontal"* problem-solving integration, here we present an exercise of analytical deconstruction of the basic requirements to develop sociological scientific knowledge -and some other usual associated practices-. This is just a preliminary version, produced as a tool to improve open discussion about the mutual contribution opportunities in the context of the SOCIAL.PATH symposium.

This catalogue has been organized in a two-column format, with three main sections: Theory requirements, Methodology requirements and Intervention requirements. The hierarchical order of sociological requirements (left side) is horizontally linked with some tentative examples of AI models, procedures, methodologies, tools and statements, but most of the right side cells are left in blank to provide open opportunities to contribution and discussion.

The aim of the following ANNEX table is not to focus on discussions about the left-column analytical descriptions, but to walk towards establishing conceptual and practical links between scientific communities. As an example taken from the table: if

[1] Dept. of Sociology, Autònoma University of Barcelona, Spain. Email: Miguel.Quesada@uab.cat.

some sociology practitioners understands the individual action as a body response for the realization of previous decisions or heuristics, and -in the later case- *"intuition"* is understand as adaptive capacities of quick decision, or unconscious motivation heuristics (Cosmides, Gigerenzer) in line with the idea of reproduction of social practices (Bourdieu) or conspicuous leisure (Veblen), then a link could be established with a number of issues and current realizations in the AI domain, such as *"embodied agents"* -with sensor-motor skills relevant to higher reasoning-, *"artificial neural networks"* -to simulate structures inside the brain that give rise to heuristic skills-, *"statistical approaches to AI"* -that mimic the probabilistic nature of the human ability to guess-, or *"reactive planning"* -for situated AI systems-.

7 CONCLUSIONS & FUTURE WORK

Most of the work presented here has focused on what we consider the two main requirements to develop sociology as a science, (a) explanatory foundations from lower ontological levels -not social, but biological and psychological-, and (b) basic epistemological and observational issues.

Future work will be oriented to complete the table with: (1) an extended revision of sociological literature in order to better capture the explanatory low-level foundations, (2) a more in-depth specification of applied theories to specific social phenomena [6], and especially (3) improvements in the links between AI and sociological research, as result of the SOCIAL.PATH conference discussions and other similar initiatives.

REFERENCES

[1] H. A. Simon. The Sciences of the Artificial. The MIT Press, Cambridge, MA (1996).

[2] F. Squazzoni. Agent-based computational sociology. Wiley & Sons, Hoboken, N.J. (2012).

[3] R. K. Sawyer. Social emergence : Societies as complex systems. Cambridge University Press, Cambridge, N.Y. (2005).

[4] F. J. Miguel. Acceptability and Complexity: Social Scientist Dilemma, and a ABSS Methodology case example. *Inteligencia Artificial*, 13(42): 21-35 (2009).

[5] R. Axelrod. Agent-based Modelling as a Bridge Between Disciplines. In K. L. Judd & L. Tesfatsion (eds.), Handbook of Computational Economics, Vol.2: Agent-Based Computational Economics. Handbooks in Economics Series, North-Holland (2005).

[6] P. Hedström and P. Bearman. The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford, UK (2011).

ANNEX: A CATALOGUE OF SOCIOLOGICAL SCIENCE REQUIREMENTS AND AI LINKS

SOCIOLOGY REQUIREMENTS	A.I. ACHIEVEMENTS
THEORY	
1. Explanatory foundations, from lower ontological levels (bottom-up explanation).	
A) Physics and Biology: evolutionary emergence and survival of individuals	
a.1. "Fundamental Constrictions" (kind of selection?), in the inert world (existence, stability)	- Modeling of physical processes, such as constraints of social processes. - Environmental representation, and dynamics. Agent-Environment interaction.
a.2. "Natural Selection" (adaptability) in the life world (mutation, selection).	- Optimization algorithms, like evolutionary algorithms. (mutation & crossover & survival). - *"Bottom-Up"* approach: to rely on elementary behaviour, which can be combined to implement more complex behaviour. - *"Behavior-Based"* approach: not to rely on a symbolic description of the environment, but rather on a model of the interactions of the entities with their environment.
a.3. "Cultural Selection" (creativity) in the cultural world (tools, arts, religions, science).	- *"Artificial intuition"* and *"Artificial imagination"*, systems with a neural architecture like Thaler's *"Creativity Machine"*.
B) Psychology	
b.1. Emotions: Affective fundamental energy that expresses the motivations, and reach consciousness as "desires". - Background Emotions (energy, enthusiasm, calm, anxiety, mood), - Elementary Emotions (fear, anger, disgust, surprise, happiness, sadness) - Secondary / Social Emotions (jealousy, love, hatred, guilt, gratitude, sympathy, embarrassment, shame, envy, anger, contempt, pride).	- *"Multi-agent systems"* with communication protocols and learning algorithms.
b.2. Motivations: bodily or cognitive homoeostatic responses generated and managed at different neuronal system locus.	
1) Selfish: Generation of personal first-order gain.	- *"Intelligent agents"*, as a system that perceives its environment and takes actions that maximize its chances of success.
1.1. Autonomy / Self-determination: The experience of being causative agent of actions.	- AI entities, autonomous in their environment, thanks to both the intrinsic robustness of the control architecture, and its adaptation capabilities to unforeseen situations.
- Instrumental Motivations: Oriented by the satisfaction derived from the consequences of the action.	
- Expressive Motivations: Oriented by the satisfaction of the execution of the action.	
1.2. Control / Personal competence: Influencing environmental conditions.	
- Power Motivations: Control over the actions of other agents (=> "dominance relations").	

- Achievement Motivations: Personal competence for success in evaluated activities (=> *“market relations”*).	
1.3. Opportunistic selfishness Motivations: Cooperate with other by the expectation of immediately recovering the cost involved.	
2) Altruistic: Generation of improvement for others' personal wellbeing.	
2.1. Kin Motivation: *“biological altruism”*, with genetic foundations.	
2.2. Strong (unconditional) altruistic Motivations: Cooperate with others and punish those who violate rules of cooperation, with no expectation of recovering the cost involved.	
2.3. Conditional altruism Motivations (reciprocity): Cooperate with others and punish those who violate rules of cooperation, with expectation of recovering later the cost involved.	
b.3. Intentional consciousness: Physiological process that represents the own mental states -motivations, emotions, perceptions, beliefs, memory- or the other's ones -up to 6th grade- (*“Inner Eye”*, Humphrey, 1986).	
1) Beliefs: Representations of the state and operation of the world -current, past or future-.	
- Factual Beliefs: Representations of the state of the world, the mental content of the others, and the fitting between means and objectives.	- Knowledge representation, Ontologies and other Top-down approaches. - Robotics sub-problems, like *“localization”, “environment mapping”* .
- Normative Beliefs: Evaluation on the adequacy of self or others' beliefs, desires or actions.	
2) Desires: Representations of the self motivations for action.	
3) Feelings: Representations of the self emotions associated with an action.	
4) Decisions: Plans or Intentions (Bratman, 1999) to future actions.	- *“Markov Decision Processes”*, Optimization planning. - *“Multiagent planning”*, and other emergent behavior approaches that uses the cooperation and competition of many agents to achieve a given goal (*“Adversarial Planning”, “Advanced Planning”*).
b. 4. Formation of intentional consciousness:	
1) Beliefs formation	- *“Hybrid intelligent system”* with both symbolic, sub-symbolic and reactive components.
1.1. Learning: Integrating the environmental captured information.	- *“Situated or behavioral AI”* with systems that behave realistically in their environment.
- Individual: By means of trial, error, reward and reinforcement.	- *“Machine learning”* with algorithms as “reinforcement learning” (where correct input/output pairs are never presented, nor sub-optimal actions explicitly corrected) or “supervised learning” (with training data criteria).
- Social: By means of imitation, by means of communication.	- *“Machine perception”, “Computer vision”* and other pattern matching classifiers methods. - *“Affective computing”* with systems that can recognize, interpret, process, and simulate human affects. - Symbolic language processing.

1.2. Abstract Intelligence: Capturing certain aspects common to many instances or environmental phenomena, although different in other respects (abduction and induction)	- *"Neural net"* approach to simulate the structures inside the brain that give rise to heuristic skills. - *"Default reasoning", "Ontological engineering"* and *"Commonsense knowledge"* (logic-based, or *"scruffy"*, or sub-symbolic).
1.3. False beliefs:	
- Individual formation: Cognitive difficulties and biases ("cold mechanism", Elster), or Interaction with emotions and desires ("hot mechanism", Elster)	
- Social formation: By means of rumor diffusion.	
2) Desires formation:	
2.1. Evolutionary structures and biological needs.	
2.2. Constraints of opportunities.	
2.3. Frustrations, or feelings of lack.	
2.4. Symbolic associations with socially desirable attributes.	
b.5. <u>Individual Action</u>: Body response as realization of decisions or heuristics.	- The sociological *"atomic unit"* is action, so no strong-AI commitment: "*The primary mission of artificial intelligence research is only to create useful systems that act intelligently, and it does not matter if the intelligence is "merely" a simulation"* (Russell & Norvig 2003, p. 947)
1) Intuition: Adaptive capacities of quick decision, or unconscious motivation heuristics [Cosmides, Gigerenzer]. Reproduction of social practices (Bourdieu), or conspicuous leisure (Veblen)	- *"Embodied agents"* with sensor-motor skills relevant to higher reasoning. - *"Artificial neural networks"* approach to simulate structures inside the brain that give rise to heuristic skills. - *"Statistical approaches to AI"* the mimic the probabilistic nature of the human ability to guess. - *"Reactive planning"* of situated AI systems.
2) Instrumental rationality: Process of evaluation and optimization of available resources -the optimal beliefs- in order to achieve a goal -desire- preset by motivations -selfish or altruistic-.	- *"Automated reasoning"*, with different forms of logic (propositional, first-order, fuzzy, subjective) - Efficient problem-solving algorithms, modeled as step-by-step deduction with *"cognitive simulation"* approach (early AI research, GOFAI, the SOAR architecture) - Action selection mechanisms, modeled as *"subsumption architectures", "free-flow hierarchies"* and *"activation networks"*.
2.1. Economic rationality: Consider just the selfish motivations.	- *"Game theory"* approach, with utility-based models ("Markov decision processes" or "dynamic decision networks / trees").
2.2. Bounded rationality: Requires just goal satisfaction, not optimization (Simon), or requires "good reasons" for the action (Boudon).	
2.3. Non-consequentialist rationality: Requires applying rationality also to the generation of goals (Searle, Boudon, Sen).	
3) Divergence between decision and action	
3.1. Akrasia (Davidson): voluntarily taking an alternative action.	
3.2. Weakness of will (Holton): unjustified revision of previous rational evaluation.	
3.3. Gap (Searle): justified revision, by 2nd order altruistic or moral evaluation, so modifying the preferences order.	
3.4. Social influence: Imitative adapting to the social modal behavior -e.g, informational cascades-.	

C) Anthropology, and other Social Sciences.	
c.1. Elementary forms of sociability (Fiske): co-evolutionary human brain structures that offer four basic relational models (=action coordination partner selection ACPS), combined differently to generate any specific culture configuration:	- *"Situated AI bottom-up"* approach: to rely on elementary behaviors, which can be combined to implement more complex behaviors.
1) Community Sharing: Equivalence between those recognized as equals (parochiality)	
2) Authority Rating: Asymmetry between those recognized as hierarchically ordered (authority)	
3) Equality Matching: Direct reciprocity among participants, with reputation memory (reciprocity)	
4) Market Pricing: Socially significant proportionality through utility units (market).	
c.2. Shared specific cultural patterns: Individuals cluster recognition.	
1) Values: affective predispositions, socially reinforced, that guide actions.	
2) Beliefs: States of the world, including mental representations of beliefs of others.	
3) Norms: Patterns of behavior resulting from actions -based on shared values and beliefs- oriented to avoid emotional social dissonance (shame, guilt, punishment).	
4) Coordination equilibrium: aggregate behavior patterns, based on practical interest (conventional).	
c.3. Social Organization / Structure: Specific environmental and social resources access distributions.	- Biologically inspired algorithms such as *"swarm intelligence"*.
1) Kin groups: Derivatives of the organization and control of human reproduction (demographic model).	- *"Evolutionary computation"* that mimics the population-based sexual evolution through reproduction of generations.
2) Social Networks: Interpersonal coordination for accessing to resources of production activities (social models or network topology).	- Communication networks
3) Functional division of work / Social Inequality: Generation and strengthening of functional specializations related to unequal rewards (dominance model).	- *"Task allocation"* algorithms
2. Applied theory to specific phenomena	
A) Social selection (agents interaction)	
a.1. Intentional posture a.2. Observer's point of view a.3. Increasing survival capacity a.4. Increasing performances a.5. Conflict resolution	
B) ...	
METHODOLOGY	
1. Epistemology	
A) Simplify the complexity of social ontology, to achieve locally useful knowledge.	- *"The Architecture of Complexity / Near Decomposability of Social Systems"* (Simon, 1962, 1996)

B) Causal attribution of social phenomena, not to another social phenomena, but to a lower-level phenomena.	- Understanding the user's psychology, cognitive behavior and problem-solving patterns (Simon, 1996) - *"DAI Rational agents"* as systems with distributed architectures and emergent properties.
C) Statistical generalization is less-informative, compared with causal generalization and the use of **explanations** through **causal mechanisms** (not laws).	
c.1. Causal mechanism is a pattern that occurs frequently, that is easily identifiable, and that is triggered under certain conditions with some consequences determined by the context of causal interaction.	- Probabilistic methods for uncertain reasoning, like *"dynamic Bayesian networks"*. - Multi-agent based systems, with *"situated"* productions rules. - *"Activation functions"* similar to that in artificial neural networks.
2. Observation / Data Generation	- *"if the problems relate to physical objects, they (or their solutions) can be represented by floor plans, engineering drawings, renderings, or three-dimensional models. Problems that have to do with actions can be attacked with flow charts and programs."* (Simon, 1996)
A) Objective behaviors: Sensitive to the observing practices.	
B) Subjective mental states: unobservable, and under risk of overinterpretation ("theory of mind").	- *"States of artificial minds"* (Ferber, 1999)
3. Validation / Data Analysis	
INTERVENTION, AND SOCIOLOGICAL NON-SCIENTIFIC PRACTICES	
1. Descriptions, useful for making founded decisions.	
2. Founded essay, based on the evolution of society, that provides an overview and draws implications for political or social action, without strong validation hypotheses (*"Philosophy of History"*).	
3. Critically oriented objective knowledge or social intervention, to put on the political or social agenda some issues regarding social minorities.	
4. Literature and social films, with thick descriptions of behaviors, processes and social phenomena.	

Table 2. Tabulation of some Sociology requirements and Artificial Intelligence achievements.

Modelling Normative Awareness: First Considerations

Paul Rauwolf[1] and **Tina Balke**[2] and **Marina De Vos**[3]

Abstract.
As software agents are being employed in more complex situations, experimental findings in the social sciences are becoming increasingly relevant to the computational sciences. The social scientific concept of situation awareness is now being utilized to quantify the success of an agent's environmental perceptual comprehension and causative processing. In this paper, we suggest that awareness of one's situation is not sufficient to succeed in navigating the growing complexity of agent-based social interactions. In human societies norms (personal, legal, and social) have emerged as multi-faceted mechanisms for prescriptive pressures projected onto individual's beliefs and intentionality. Here we define the term normative awareness as the perceptual comprehension of norms and the prediction of the causative effect of actions on norms. In this paper, we suggest that such awareness of the creation and perpetuation of norms would prove advantageous to agent-based research and review to what extent the multi-agent system literature has implicitly utilized the concept of normative awareness. We recognize that a ubiquitous merger of the vernacular between the social and computational sciences is unnecessary, as such, we discuss when and how normative awareness should be extended to agent-based modelling and multi-agent systems.

1 A Case for Modelling Normative Awareness

Situation awareness, which includes perceptual processing, comprehension, and causative predictions [23], is a foundational skill in generating useful human action selection mechanisms. Recently, this concept has been projected onto the study of computational agents (see [42, 30, 31] for example), permitting quantified measurements of awareness, and thus opening a dialogue of the utility therein. However, the instantiation of multiple agents within the computational arena may lead to further complexities than those described in the situational awareness literature. Thus, we define a new term, *normative awareness* as projected onto the situation awareness definition, as the perceptual processing, comprehension, and causative predictions of norms.

The social interactions inherent in multi-agent systems generate normative complexities analogous to those described in the social sciences. Rather than perceiving a situation and processing the potential consequences of an action in isolation, agent's performance is increased through deftly navigating social nuances. Historically, humans have employed norms (personal, social and legal) in manoeuvring the complexities of social interactions. Successful navigation has thus been aided by awareness of the propensity of others to ascribe to such norms, as well as a prediction of the beliefs and intentionality of others. It has even been argued that avoiding sanctions, weeding out defectors, and the emergence of cooperation could be linked to society's tendency to instantiate and navigate norms[35].

We suggest, that as situation awareness is being employed in the agent-based literature, so too should normative awareness. We argue that agent awareness of normative underpinnings are sufficiently unique to situational awareness as to warrant a separate definition. Rather than simply processing environmental data, normative awareness employs a limited version of *Theory of Mind* [7], in that agent's are aware that other agent's possess awareness, intentionality, and beliefs. We argue that this definition transcends that of situation awareness, and that a dialogue surrounding the benefits of normatively aware agents will prove useful in grounding future agent-based research.

To justify this postulation, in Section 2 we first present a brief introduction into the social science and computational literature regarding situation awareness. Next, in Section 3, we define normative awareness in reference to situation awareness. We approach this in a multi-faceted way. In characterizing normative awareness, the semantic ideology of norms is considered, since awareness of norms begs a definition of norms. The definition, utility, creation, and perpetuation of norms (within the social sciences) are thus inspected, and while debate continues, a broad spectrum of arguments are considered. The goal is not to cement a rigid criteria for norms, but rather to discuss the breadth of research in order to (i) augment the grounding of future computational instantiations of norms in theory, and (ii) aid in unifying the vernacular between the two disciplines. Upon exploring norms and situation awareness, we propose a definition of normative awareness, which amalgamates the two concepts. Section 4 reviews the existing agent-based literature regarding normative awareness. We also discuss the gaps in usage between the social sciences and the computational sciences, and discuss whether these gaps need to be closed or not. Finally, in Section 5 we conclude our discussion with a summary and an agenda for future research.

2 Situation Awareness

Situation awareness (SA) has a history of use within military aviation vernacular, dating back to World War I [24]. More recently the term has been utilized within the social [40] and computational sciences [30, 42]. As this paper juxtaposes normative awareness to situation awareness, in this section we will discuss the idea of SA in more detail. We in particular focus on the work of Mica Endsley, who is well known for her work on SA and who defines SA as follows:

> *Situation Awareness* - perception of the elements of the environment within the volume of time and space, the comprehension of their meaning, and the projection of their status in the near future [23].

[1] University of Bath, UK, email: p.rauwolf@bath.ac.uk
[2] CRESS, University of Surrey, UK, email: t.balke@surrey.ac.uk
[3] University of Bath, UK, email: mdv@cs.bath.ac.uk

Based on this definition, Endsley develops a three-layer hierarchical structure which is often referenced when discussing SA [24]:

Level 1: perception of the elements in the environment. This is the identification of the key elements or "events" that, in combination, serve to define the situation. This level tags key elements of the situation semantically for higher levels of abstraction in subsequent processing.

Level 2: comprehension of the current situation. This is the combination of level 1 events into a comprehensive holistic pattern, or tactical situation. This level serves to define the current status in operationally relevant terms in support of rapid decision making and action.

Level 3: projection of future status. This is the projection of the current situation into the future in an attempt to predict the evolution of the tactical situation. This level supports short-term planning and option evaluation when time permits.

Endsley's hierarchical nature of the SA theory has lent itself well to the computational sciences, permitting the awareness of an agent to be discussed in a grounded way. The notion has been utilized in coordinating agents operating within service based systems [42], as well as formally quantifying the awareness of an agent via its ability to complete truth tables in a particular context [30]. Additionally, situation awareness has been employed in the creation of military tactical plans through agent-based modelling [31].

However, as much as increasing SA augments predictive ability for potential actions, it does not offer a holistic theory to guide action selection mechanisms. When one is highly situationally aware, one accurately perceives and comprehends the environmental context. This permits veracious projection of the consequences of a given action. But, what action will be selected? What is one's motive? Social theorist Paul Stern argues that situation awareness is a necessary but not sufficient condition for social movement [39]. He suggests that the motivational impetus to act often portrays itself as a sense of obligation, or a personal norm. Thus, to Stern, social movement requires (i) the ability to predict the future outcomes of actions utilizing situation awareness, and (ii) awareness of which actions and direction one wishes to push society (i.e. awareness of one's personal norms and goals). This suggests that awareness of norms adds an additional layer to interpreting a situation juxtaposed to a solely situationally aware agent. Furthermore it requires an understanding of the link between the SA awareness of the agent and the interpretation of this situation (e.g.action, observations,... with respect to the norms of the society.)

This presence of norms as motivational factors in human and multi-agent societies may complicate the notion of awareness. Should an agent be aware of normative societal underpinnings? If the awareness of norms enhances the ability to project future status, does such a notion fall under the banner of situational awareness? In the next section we discuss the definition of norms, and in that attempt to diagnose the utility of diverging the definitions of situational and normative awareness within multi-agent systems.

3 Normative Awareness in the Social Sciences

3.1 Definition of Normative Awareness

If Endsley's situation awareness theory is projected onto norms, then normative awareness is the *(i) perception of norms, (ii) comprehension of norms, and (iii) ability to predict future system states given norms.* However, this definition is unsatisfactory without semantic cohesion. What is a norm? Can awareness of norms be implemented as a subset of situation awareness? Is a norm a situation?

To answer these, first, we will describe the breadth of the social scientific usage of the term "norm". Next, we will briefly discuss theories on the creation and perpetuation of norms. In the end we will present an argument that while Level 1 and 2 situation and normative awareness may prove indistinct, it is only in understanding an agent's effect on the norm that an agent will attain level three awareness, future projection. It is postulated that level three normative awareness requires at least a limited version of theory of mind, in that the agent must predict motivation's and actions of other agent's based on their goals and beliefs.

3.2 Definition of Norm

Before delving into the nuances of normative literature, it may prove useful to reiterate our intention. In discussing social science's utilization of the term norm, our goal is not to reach a conclusion on a definition still debated or to get mired in a semantic argument. Rather, in articulating the breadth of the terminology it may yield the knowledge requisite to deliberate upon the utility of passing aspects of the vernacular into the computational arena. Additionally, in acknowledging the historic debate and precedent, multi-agent systems can ground itself in existing theory.

In a general sense, the social science literature considers norms to be prescriptive and proscriptive [9]. There are actions which ought to be employed and actions which ought not. This pressure may be placed on the self, in which case it is considered a personal norm. In contrast, social norms are rules that are:

> neither promulgated by an official source, such as a court or legislature, nor enforced by the threat of legal sanctions, yet [are] regularly complied with (otherwise it would not be a rule) [36].

Some, however, argue that this notion of the burden of enforcement leads to an even further refined differentiation in the nomenclature. The term "convention" has been employed to describe a Nash equilibrium of a cooperative game [34]. Though there may be multiple equilibria, once convergence reaches a certain threshold, it is rarely in one's interest to defect. For instance, walking on the "incorrect" side of a footpath seldom requires social sanctions as the defective act is cost prohibitive. In general, conventions "...provide people with means of knowing what to expect of each other and thereby serving to coordinate interactions [41]."

Bicchieri [9] argues a social norm is a mechanism which alters a mixed-motive game[4] into a cooperation game. For instance, normative prescriptions and the potential for sanctions might alter the cost/benefit utilities of a context analogous to a prisoner's dilemma game [5] (where cooperation is not a Nash equilibrium[5]) into a situation where cooperation is a Nash equilibrium. Such instances typically require sanctions in order to manipulate the topology of utility function. For example, the utility of attempting to steal is altered depending on the consequences of being caught.

While the differentiation in the governance of the norms has led to a distinction between personal, social, and legal norms, these definitions still blur [9]. Stern argues that personal (rather than social)

[4] A mixed-motive game consists in a game where the best pay-off for at least one of the players does not lead to the best pay-off for the other.

[5] A Nash equilibrium is a state in game theory where, if all other players do not alter their action, it is not beneficial for any one player to alter their action.

norms are required to change the social landscape precisely because the status quo being overthrown typically involves a relatively ubiquitous social norm [40]. The delineation proves a bit more semantic when considering whether a personalized social norm is both a personal and a social norm, or a social norm enacted through an individual.

Additionally, the definition and delineation between legal and social norms presented above are under dispute. Conte and Castelfranchi argue that colloquially it is accepted that behaviour is not sufficient for defining accepted prescriptive pressures. They suggest that just because people throw their rubbish out the window does not suggest that people ought to throw their rubbish out the window [15]. However, if normative prescription does not alter behaviour and sanctions, then how could it alter mixed-motive situations? Furthermore, Fehr and Fischbacher argue that legal norms only exist as a epiphenomena of social norms.

> Legal enforcement mechanisms cannot function unless they are based on a broad consensus about the normative legitimacy of the rules – in other words, unless the rules are backed by social norms. Moreover, the very existence of legal enforcement institutions is itself a product of prior norms about what constitutes appropriate behaviour [27].

This need for broad consensus has even lead to some paradoxical interplay between laws and social norms. In Alabama, a law outlawing adultery was not repealed due to the presumed political difficulty in passing the legislation, despite strong sentiment against enforcement [29]. While a law, is it a legal norm or social norm? If law is not legislatively enforced, should a person or agent care? In this example, was law being employed as a mechanism for disseminating social norms even without the threat of legal sanctions? Is the law there as a referent to the possibility of social sanctions? Is there an expectation or a utility in being aware of the legislation?

3.3 Creation of Norms

In discussing the definition of norms it may help to acknowledge the debate regarding normative generation. How and when are norms created? Is there some tipping point in terms of percentage of the population that generates enough pressure to alter the social landscape? When does societal preference lead to societal pressure?

Bicchieri [8] suggests that the creation of norms are analogous to the formation of language. Namely, she argues that the prescriptive and proscriptive pressures underlying normative interaction are similar to the grammatical structure in language in that neither were the result of human planning, but rather emerged. Conte and Castelfranchi [15] posit that, while norms are spoken of definitively, normative pressure and thus the existence of a norm lies on a continuum consisting of how pervasively people (i) behaviourally conform, (ii) believe they should conform, (iii) are spatially distributed.

Additionally, they discuss the differentiation in the literature between the epiphenomenal and evolutionary generation of norms. While the epiphenomenal explanation relates to the game theoretic conversation, the evolutionary approach argues that norms are generated due to bounded rationality. By implementing prescriptive mechanisms agents can limit the need to diagnose other agent intentionality as well as comprehending the repercussions of complex social interaction [15]. It is even been argued that the advent of social norms offers the advantage of reducing the need to think [25].

As previously mentioned, Fehr and Fischbacher suggest that legal norms are only begotten through the consequences of social norms. However, the interaction between social and legal norm creation are recursive. Scott [38] notes that once a legal norm is established society may adopt it as a social norm even if the legal establishment does not provide it is typical sanctions for defection. For example, a "no smoking" sign may generate social sanctions (e.g. shaming) even if the governance which placed the sign does not police the policy.

3.4 Perpetuation of Norms

Lastly, a definition of norms should consider the perpetuation and declination of norms. When is a norm no longer a norm? Legal positivism argues that the validity of a law's existence need not require general social consent, rather if the authority responsible for legislation pens a new law, it is a law. This is in contrast to the argument that laws may be abrogated due to desuetude (i.e. disuse, or not being enforced) [29]. This ideological disparity becomes relevant when considering multi-agent systems, and whether defecting against a priori norms (even if desuetude) always constitute a violation, and thus advocates awareness. Thus far, there is a propensity in mult-agent systems field to invoke legal positivism, in that defection is always "illegal" [17].

More definitively, a social norm ceases to exist if no one expects anyone else to employ it. Thus, the consequences to a norm given an agent's action depends on the type of norm, and potentially (i.e. in the case of legal norms) the ideology grounding the norm. If an agent defects against a legal norm by the legal positivist definition, the norm remains unaffected. Conversely, if an agent defects against a social norm, the strength of the norm is affected [15].

Additionally, if agent behaviour, at least in part, defines and perpetuates norms, then it is not solely the decision to defect which alters the strength of the norm, but also the agent's propensity to sanction defectors. Even further, an agent's tendency to sanction agent's who refuse to sanction may aid in perpetuating the norm [4].

Thus, at least in the case of personal, social, and legal (given abrogation via desuetude) norms, perpetuation is dependant upon agent belief and, arguably, action. Therefore, awareness of norms is enhanced via awareness of one's and other's beliefs.

3.5 Leading toward Normative Awareness

Having briefly noted the breadth of the literature regarding norms, it is more feasible to discuss the implications of normative awareness. First, it is necessary to decide whether there is any difference between situation and normative awareness. If such a valuable distinction is uncovered, then the utility of normative awareness may be debated.

We suggest that there is little value in differentiating the two forms of awareness at the first two levels. An agent who is level 1 and 2 normatively aware is essentially situationally aware. To perceive and comprehend normative prescriptions is not usefully distinct from perceiving and comprehending environmental situations in that perception and comprehension of norms *is* an environmental situation. However, we believe that the concepts diverge at level three awareness, such that level three normative awareness constitutes level three situation awareness, plus awareness of personal and other agent's motivations and normative restrictions.

As previously mentioned, Endsley's third level of situation awareness elucidates the consequences of actions, but does not consider which action to select. On an individual level, Stern suggests that social movement not only requires situation awareness, but also awareness of one's personal norms [40]. Thus, at a personal level, third level normative awareness includes not only the ability to predict, but also the knowledge that one wishes to act. In other words, awareness of one's personal norms offers more information than simply the potential consequences of taking an action, it suggests what action will be taken. High situation awareness garners accurate predictions of consequences given an action, while personal normative awareness also posits which actions could be taken.

Furthermore, at a social normative level (or legal normative level sans legal positivism) it has been argued that the perpetuation of a given norm must be dependent, to some extent, on agent behaviour. As such, if one's action effects the perpetuation of a norm, it also effects the creation of a norm, even if the new norm is not perpetuating the old norm (i.e. new norm B is not acting upon or sanctioning old norm A). Thus, for normative awareness to reach level three situational awareness (i.e. projection of future status), then an agent must be aware of other agent's ability to affect norms, and thus there is utility in the cognizance of motivation and goals.

It could potentially still be argued that normative awareness is a subset of situation awareness, in that awareness of the environment includes awareness of other people's motivations, beliefs, and potential for sanctioning. Why confound the terminology when one could potentially extend the definition of situation awareness to include beliefs, etc? We suggest that, even if normative awareness is reduced to a subset of situation awareness, the nuanced and complex effects of beliefs will present unique problems. Even if one remains unconvinced in the semantic delineation between normative and situation awareness, evolution seems to have handled the differentiation through unique neurological processes. Leda Cosmides demonstrated that we are better able to draw logical inferences when the data is set in social rather than mathematical contexts [18]. Analogously, when syntactically instantiating multi-agent models, the processing of beliefs and the prediction of other agent's beliefs are typically unique modules compared to the algorithms employed for situation awareness. In other words, the programmed modules for situation awareness, and processing other agent's beliefs and motivations will likely prove different modules. Thus, even if semantically the concepts can be amalgamated, practically they may be programmed separately, which then creates utility in semantically differentiating the algorithms which will prove conceptually distinct.

3.6 Utility of Normative Awareness

From a utility perspective, advanced normative awareness is likely beneficial. Although, in certain circumstances this can be argued. If legal norms exist, but are desuetude, then awareness may prove deleterious from the standpoint of cognitive load.

Furthermore, in human society lacking normative awareness can sometimes protect one from sanctions. Children, and the mentally disabled are often given a reprieve from the social effects of defection given ignorance. This has raised philosophical debates regarding the norm of sanctioning, including whether psychopaths should be punished if they can not differentiate between conventions and morality [33], or whether one can avoid social and legal sanctions by claiming emotional distress (e.g. temporary insanity)?

While these arguments are potentially rare and nuanced occurrences in human society, philosophically they are useful in discussing normative awareness in the computational arena. Laws, which are always human constructs in the real world, are not always agent constructs in the agent world - they may be designed by humans. If all agents defect from a law, is it useful to be aware of it? Additionally, awareness of one's capacity for awareness has been integrated into a human understanding of norms, but what about within multi-agent systems? If agents are homogeneous in their capability, then perhaps it is a non-issue. But, how should agents be developed when one agent is capable of a deeper normative awareness than another? Should the more advanced agent sanction the other even though the agent will never comprehend the situation? If not, should a capable agent pretend ignorance? Would such a situation ever prove useful?

4 Computation Models of Normative Awareness

Having discussed the concept of norms and normative awareness and having projected it onto the situational awareness levels by Endsley, in this section we now shift our focus to the computational modelling of normative awareness. For this purpose, we start by reviewing the literature on current (computational) models of normative agents[6].

4.1 Literature Review

Turning to the normative agent architectures first, the most prominent ones found in the literature are BOID [11], NoA [32], BRIDGE [20], deliberative normative agents [12], EMIL-A [2] and the NBDI architecture [19, 22].

Of these different frameworks, BOID does consider agents reasoning about norms, but it is assumed that all norms are known to the agents. Normative awareness as such is therefore not considered, but what we refer to as Level 1 and 2 normative awareness is automatically assumed. This is similarly true for NoA and BRIDGE. Although Dignum et al. state that "A person may be aware of a norm . . . ", but do not explore the issue further.

The idea of deliberate normative agents is based on earlier works in cognitive science (e.g. [15, 16]). Similar to BOID it focuses on the idea that social norms need to be accounted for in the decision making process of an agent.

As a result of the complexity of the tasks associated with social norms, [12] argue that they cannot simply be implicitly represented as constraints or external fixed rules in an agent architecture, but they suggest that norms should also be represented as mental objects, which have their own mental representation [14] and should interact in several ways with the other mental objects (e.g. beliefs and desires) and plans of an agent. Looking at the generation of these mental objects, they result from an internalization process by the agents that is briefly outlined in [12]. For the internalization, when agents are in a social setting, norms are immediately recognized as such (either by observation or communication) and agents can determine to internalize them, i.e. to incorporate them into their own decision making or not (depending on their attitude towards the specific norms and its consequences). This decision making mainly incorporates the ideas of Level 1 and 2 normative awareness, i.e. the agent – as in the previously mentioned architectures – focuses on the question to what extend the norms will affect its own behaviour, but does not necessarily consider its effect on other agents.

The idea of the internalization process described in [12], as well as the actual recognition of norms as such is extended in the EMIL-A architecture [2, 3], which uses a specific norm recognizer module for the latter. This module distinguishes two different scenarios: (i) information it knows about and has classified as a norm before and (ii) new (so far unknown) normative information.

[6] In our review of models of normative agents, except for [6] we neglect models focusing on designing normative frameworks such as the eInstitutions [26, 28], InstAL [13], OperA [21], $\mathcal{MOISE}^{Ins}$ [10],. . . as these tend to focus on the normative architecture, rather than the agents and their reasoning.

In the former case, i.e. the agent receiving an external normative input it is already aware of, the normative input is entrenched on a so-called normative board (which captures the long-term and working memory of an agent) where it is ordered by salience. Here, salience [1] refers to the degree of activation of a norm, i.e. how often the respective norm has been used by the agent for action decisions or how often it has been invoked. The norms stored on the normative board are then considered in the classical BDI decision process as restrictions on the goals and intentions of the agent. In this process the salience of a norms is important, as in case of conflict (i.e. several norms applying to the same situation), the norm with the highest salience level is chosen.

In case the external normative input is new, i.e. not previously known to the agent, the agent needs to internalize it first before being able to apply it in any of its decision making. For this purpose the normative frame is activated. The normative frame is equipped with a dynamic schema (a frame of reference) which is used to recognize and categorize an external input as being normative based on its properties. Properties that the normative frame takes into account are for example deontic specifications, information about the locus from which the norm emanates or information about legitimate reactions or sanctions to transgression of the norm[7]. The recognition of a norm by an agent does not necessarily imply that the agent will agree with the norm or that it understands it fully, it only means that the agent has classified the new information as a norm. After this initial recognition of the external input as a norm, the normative frame is used to find an interpretation of the new norm. This is done by checking the agent's knowledge for information about variables of the frame of reference for example. Once enough information is gathered about the new norm and the agent is able to determine its meaning and implication, the newly recognized norm is turned into a normative belief. Again, normative beliefs by an agent do not imply that the agent will follow the respective norm, instead it is a candidate for a norm that the agent might adopt. With respect to the adoption of norms, in EMIL-A agents follow a "why not" approach. This means that an agent has a slight preference to adopt a new norm if it cannot find evidence that this new norm conflicts with its existing mental objects. Adopted normative beliefs are stored as normative goals. These normative goals are considered in the agent's decision making. An agent does not need to follow all normative goals it has when making a decision, but can violate norms. When deciding whether to follow a norm, EMIL-A assumes that an agent will try to conform with its normative goals if it does not have reasons for not doing so. As in the previous agent architectures, level 1 and 2 only.

A different approach to the one of EMIL-A concerning the identification of emergent norms (levels 1 and 2) without the norm explicitly being given to the agent is presented by Savarimuthu et al. [37]. The authors present an approach to use data mining mechanisms to identify prohibition and obligation norms, however stop after the norm identification, i.e. they do not account for level 3 either.

One of the few normative architectures which allows for incorporating considerations on how other agents react to norms (i.e. the requirement for level 3 normative awareness) is presented in [6]. In this architecture, using queries, (via an intermediary) the agents can ask about the norms of the system (and in particular the ones applying to them) as well as pose queries about the effects of their own actions or the actions of others with respect to a desired outcome. One example of a query presented in the paper is whether a particular state can be reached depending on the actions of others if the agent itself performs a particular (sequence of) action(s). In contrast to the agent architectures described earlier, the work presented in [6] has its main focus on the normative framework, rather then the internal reasoning of agents.

As a result, the norms the agents deal with are indirectly assumed to be predefined legal norms. Social norms that are emerging in the course of the interaction of the agents are not considered. With respect to obtaining normative information, the agents in the system need to make a conscious decision to ask for normative information, no information is passed on to the agents without a query action initiated by them. The authors point out that the information gathered by their queries can be adopted by the agents in the form of percepts of the environment and then considered in the agent decision making process. This "recognition" of information via the environment is in line with the idea of perceiving a situation (of which norms are a part of) by the agents. However, this does not include any consideration as to which norms an agent internalizes and how it uses the knowledge obtained with the help of the queries in [6]. At present a general methodology or formalisation of the approach presented in [6] is missing.

Inspiration for such an agent architecture might be drawn from the NBDI agent architecture [19, 22]. This architecture includes a Norm Recognition module, which agents can use to either implicitly (via observations) or explicitly (via communication) learn about new norms. The architecture proposes bridge rules for norm internalization and considers different agent behavioural types (with respect to the adoption of norms). Again the architecture is very limited with respect to considering other agent's reactions towards norms, i.e. level 3.

4.2 Gaps and Challenges

Having reviewed existing normative (agent) architectures, it is apparent that at this stage, no model incorporating all three level of normative awareness exists. Thus, although some work has been done on norm emergence in terms of explaining how an agent decides whether it adopts a norm or not, – except for EMIL-A and the architecture described in [6] – most architectures assume that norms are automatically detected (either via observation or communication) and questions on whether something is a norm are not being asked by the agents. Furthermore, the information on norms is typically only reflected on the agents themselves and a lack of consideration of level 3 normative awareness can be found.

As pointed out before, the level 3 normative awareness indirectly implemented in [6] focuses on predefined norms specified at a systems level and thus lacks the incorporation of norms emerging via social interplay. As a consequence no need for normative awareness outside of situation awareness is required in that architecture. This focus on predefined norms is however not uncommon in the (normative) multi-agent literature. Whereas in human societies legal, social, and personal norms are human constructs, often in the computational arena norms are not defined by the agents, but they are designed into the system, which was designed mostly with a particular purpose in mind. As such the question arises on whether the specific accounting of the social norms emerging from social interactions between agents and the emergence thereof does always need to be considered when designing a computational model of normative awareness? Is it sufficient to embed normative awareness in situational awareness if only legal norms defined into a system are considered?

Even if this notion is sufficient, another question arises as a result of our literature review. Looking at the architectures considering

[7] A detailed list of properties being considered by the normative frame can be found in [1].

agents incorporating normative information in their decision making, in these architectures little information was provided on the intentionality of agents. What makes an agent actively try to perceive norms and to incorporate them into its decision-making?

Furthermore, looking at existing models of agent decision making, the models normally assume that individual agents have a fixed set of goals (desires) that they wish to achieve. This, however, makes it hard to account for the high levels of adherence to norms and cooperation found in human societies (e.g. the ultimatum game). We discussed Theory of Minds earlier on, which as one aspect implies that humans have a 'built-in' ability to understand and share the (normative) intentions of others. The incorporation of such a "we-intentionality" in agent reasoning cannot, at the moment, be found in agent architectures. Consequently – from the agents' perspective – an important prerequisite for level 3 normative awareness is missing.

Finally, there are some aspects of the emergence of norms in biological society, which, though potentially not presently relevant, may become more so as agent-based research evolves. For instance, most agents are clones, and thus, are imbued with the same capacity for informational and causative processing as their conspecifics. As such, the aspect of normative underpinnings which derived through the interactions of heterogeneity may be ignored in the computational arena. However, this may not prove a static assumption, and when heterogeneity is employed within multi-agent systems, it may prove advantageous to look to the study of the social sciences in order to ground multi-agent research.

5 Summary and Research Agenda

In this paper we discussed the definition, utility, and difficulty in modelling normative awareness for software agents. Our intention was to initiate a dialogue which considers the utility of agent awareness of the normative infrastructure, and how normative awareness differs from situation awareness. To this end, we started off by giving a definition of situation awareness based on Endsley [24] as well as looking into the concept and related properties of norms. Starting from the individual concepts, we juxtaposed the two ideas to identify where the ideas overlapped and where further considerations were required. We identified that the main difference between situation and normative awareness resided in what we referred to as level 3 normative awareness. Namely, advanced level 3 normative awareness requires a limited instantiation of Theory of Mind, in that, especially in social norms, the norms which pressure an agent's actions are affected by other agent's intentionality and awareness. Thus an awareness that other's have their own beliefs improves an agent's awareness of the norms.

When reviewing existing literature on normative multi-agent systems concerning their incorporation of normative awareness, it became evident that at present not only do most systems assume general awareness of all norms, but that level 3 normative awareness is normally not really considered. The reason for that is the lack of models of (shared) intentionality of agents, which we perceive as the most important step in our research agenda for modelling normative awareness.

A question which was raised in our analysis was whether computational models of normative awareness always need this distinction from situational awareness. Especially in cases in which norms are legal norms respecified in a system and not social norms emerging from social interaction, the perception of norms via the environment as part of the situation might be sufficient. Even if this is sufficient, current models such as [6] pursuing this approach need to adapt their focus from the normative system perspective to a more agent-centered perspective exploring in more detail how and under which circumstances an agent will generate an intention and perform an action to actively query the norms of the system it inhabits. This is therefore the second point on our research agenda: a formalisation of the work presented in [6] as well as the development of a general methodology for the above mentioned processes.

ACKNOWLEDGEMENTS

Tina Balke was partially supported by funding from the ePolicy project (European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 288147).

REFERENCES

[1] Giulia Andrighetto, Marco Campennì, Rosaria Conte, and Mario Paolucci, 'On the immergence of norms: a normative agent architecture', in *AAAI Symposium, Social and Organizational Aspects of Intelligence*, (2007).

[2] Giulia Andrighetto, Rosaria Conte, Paolo Turrini, and Mario Paolucci, 'Emergence in the loop: Simulating the two way dynamics of norm innovation', in *Normative Multi-agent Systems*, eds., Guido Boella, Leon van der Torre, and Harko Verhagen, number 07122 in Dagstuhl Seminar Proceedings, (2007).

[3] Giulia Andrighetto, Daniel Villatoro, and Rosaria Conte, 'Norm internalization in artificial societies', *AI Communications*, **23**(4), 325–339, (December 2010).

[4] Robert Axelrod, 'An evolutionary approach to norms', *American political science review*, **80**(04), 1095–1111, (1986).

[5] Robert Axelrod, 'The evolution of strategies in the iterated prisoner's dilemma', *Genetic algorithms and simulated annealing*, **3**, 32–41, (1987).

[6] Tina Balke, Marina De Vos, and Julian Padget, 'Normative run-time reasoning for institutionally-situated bdi agents', in *Coordination, Organizations, Institutions, and Norms in Agent System VII*, eds., Stephen Cranefield, M.Birna Riemsdijk, Javier Vzquez-Salceda, and Pablo Noriega, volume 7254 of *Lecture Notes in Computer Science*, 129–148, Springer Berlin Heidelberg, (2012).

[7] Simon Baron-Cohen, 'Precursors to a theory of mind: Understanding attention in others', in *Natural theories of mind: Evolution, development and simulation of everyday mindreading*, ed., Andrew Whiten, Basil Blackwell, (1991).

[8] Cristina Bicchieri, *The grammar of society: The nature and dynamics of social norms*, Cambridge University Press, 2005.

[9] Cristina Bicchieri and Ryan Muldoon, 'Social norms', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, spring 2011 edn., (2011).

[10] Olivier Boissier and Gâteau Benjamin, 'Normative multi-agent organizations: Modeling, support and control, draft version', in *Normative Multi-agent Systems*, eds., Guido Boella, Leon van der Torre, and Harko Verhagen, number 07122 in Dagstuhl Seminar Proceedings, pp. Internationales Begegnungs– und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, (2007).

[11] Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre, 'The boid architecture: conflicts between beliefs, obligations, intentions and desires', in *Proceedings of the fifth international conference on Autonomous agents*, pp. 9–16. ACM, (2001).

[12] Cristiano Castelfranchi, Frank Dignum, Catholijn M. Jonker, and Jan Treur, 'Deliberate normative agents: Principles and architecture', in *Intelligent Agents VI, Agent Theories, Architectures, and Languages (Proceedings 6th International Workshop, ATAL'99, Orlando FL, USA, July 15-17, 1999)*, eds., Nick R. Jennings and Y. Lespérance, volume 1757 of *Lecture Notes in Computer Science*, pp. 364–378. Springer, (2000).

[13] Owen Cliffe, Marina De Vos, and Julian A. Padget, 'Specifying and analysing agent-based social institutions using answer set programming', in *Coordination, Organizations, Institutions, and Norms in Multi-Agent Systems – AAMAS 2005 International Workshops on Agents, Norms and Institutions for Regulated Multi-Agent Systems, ANIREM 2005, and Organizations in Multi-Agent Systems, OOOP 2005, Utrecht, The Netherlands, July 25-26, 2005, Revised Selected*

Papers, eds., Olivier Boissier, Julian A. Padget, Virginia Dignum, Gabriela Lindemann, Eric T. Matson, Sascha Ossowski, Jaime Simão Sichman, and Javier Vázquez-Salceda, volume 3913 of *LNCS*, pp. 99–113. Springer Berlin / Heidelberg, (2006).

[14] Rosaria Conte and Cristiano Castelfranchi, *Cognitive and Social Action*, Taylor & Francis, 1995.

[15] Rosaria. Conte and Cristiano Castelfranchi, 'From conventions to prescriptions. towards an integrated view of norms', *Artificial Intelligence and Law*, **7**(4), 323–340, (1999).

[16] Rosaria Conte, Cristiano Castelfranchi, and Frank Dignum, 'Autonomous norm acceptance', in *Proceedings of the 5th International Workshop on Intelligent Agents V, Agent Theories, Architectures, and Languages*, pp. 99–112. Springer-Verlag, (1999).

[17] Rosaria Conte, Rino Falcone, and Giovanni Sartor, 'Introduction: Agents and norms: How to fill the gap?', *Artificial Intelligence and Law*, **7**(1), 1–15, (1999).

[18] Leda Cosmides, 'The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task', *Cognition*, **31**(3), 187–276, (1989).

[19] Natalia Criado, Estefania Argente, Pablo Noriega, and Vicente J. Botti, 'Towards a normative BDI architecture for norm compliance', in *Proceedings of The Multi-Agent Logics, Languages, and Organisations Federated Workshops (MALLOW 2010), Lyon, France, August 30 - September 2, 2010*, eds., Olivier Boissier, Amal El Fallah-Seghrouchni, Salima Hassas, and Nicolas Maudet, number 627 in CEUR Workshop Proceedings, pp. 65–81, (2010).

[20] Frank Dignum, Virginia Dignum, and Catholijn M. Jonker, 'Towards agents for policy making', in *Multi-Agent-Based Simulation IX*, eds., Nuno David and Jaime Simão Sichman, volume 5269 of *Lecture Notes in Computer Science*, 141–153, Springer-Verlag, (2009).

[21] Virgínia Dignum, *A model for organizational interaction: based on agents, founded in logic*, Ph.D. dissertation, Utrecht University, 2003.

[22] Baldoino F. dos S. Neto, Viviane Torres da Silva, and Carlos José Pereira de Lucena, 'NBDI: An architecture for goal-oriented normative agents', in *ICAART 2011 - Proceedings of the 3rd International Conference on Agents and Artificial Intelligence, Volume 1 - Artificial Intelligence, Rome, Italy, January 28-30, 2011*, eds., Joaquim Filipe and Ana L. N. Fred, pp. 116–125. SciTePress, (2011).

[23] Mica R. Endsley, 'Measurement of situation awareness in dynamic systems', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **37**(1), 65–84, (1995).

[24] Mica R. Endsley, 'Toward a theory of situation awareness in dynamic systems', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **37**(1), 32–64, (1995).

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

[26] Marc Esteva, *Electronic Institutions: from specification to development*, Ph.D. dissertation, Artificial Intelligence Research Institute (IIIA), 2003.

[27] Ernst Fehr and Urs Fischbacher, 'Social norms and human cooperation', *TRENDS IN COGNITIVE SCIENCES*, **8**(4), 185–190, (APR 2004).

[28] Andrés García-Camino, *Normative regulation of open multi-agent systems*, Ph.D. dissertation, Artificial Intelligence Research Institute (IIIA), Spain, 2009.

[29] Hillary Greene, 'Undead laws: The use of historically unenforced criminal statutes in non-criminal litigation', *Yale Law & Policy Review*, **16**(169), (1997).

[30] Anne-Laure Jousselme, Patrick Maupin, Christophe Garion, Laurence Cholvy, and Claire Saurel, 'Situation awareness and ability in coalitions', in *Information Fusion, 2007 10th International Conference on*, pp. 1–9. IEEE, (2007).

[31] Octavio Juarez-Espinosa and Cleotilde Gonzalez, 'Situation awareness of commanders: a cognitive model', (2004).

[32] Martin J. Kollingbaum and Timothy J. Norman, 'Norm adoption in the noa agent architecture', in *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pp. 1038–1039, New York, NY, USA, (2003). ACM.

[33] Neil Levy, 'The responsibility of the psychopath revisited', *Philosophy, Psychiatry, & Psychology*, **14**(2), 129–138, (2007).

[34] David Lewis, 'Convention: A philosophical study', (2002).

[35] Mitsuhiro Nakamura and Naoki Masuda, 'Indirect reciprocity under incomplete observation', *PLoS computational biology*, **7**(7), e1002113, (2011).

[36] Richard A. Posner, 'Social norms and the law: An economic approach', *The American Economic Review*, **87**(2), 365–369, (1997).

[37] Bastin Tony Roy Savarimuthu, Stephen Cranefield, Maryam A. Purvis, and Martin K. Purvis, 'Obligation norm identification in agent societies', *Journal of Artificial Societies and Social Simulation*, **13**(4), 3, (2010).

[38] Robert E. Scott, 'Limits of behavioral theories of law and social norms, the', *Va. L. Rev.*, **86**, 1603, (2000).

[39] Paul C. Stern, 'New environmental theories: toward a coherent theory of environmentally significant behavior', *Journal of social issues*, **56**(3), 407–424, (2002).

[40] Paul C. Stern, Thomas Dietz, Troy Abel, Gregory A. Guagnano, and Linda Kalof, 'A value-belief-norm theory of support for social movements: The case of environmentalism', *Human ecology review*, **6**(2), 81–98, (1999).

[41] Elliot Turiel, *The development of social knowledge: Morality and convention*, Cambridge University Press, 1983.

[42] Stephen S. Yau, Dazhi Huang, Haishan Gong, and Hasan Davulcu, 'Situation-awareness for adaptive coordination in service-based systems', in *Proceedings of the 29th Annual International Computer Software and Applications Conference*, pp. 107–112, (2005).

A workflow for an interdisciplinary methodology to build computational cognitive models of human social behaviours

Jordi Sabater-Mir[1]

Abstract. In this paper we present an initial workflow for an interdisciplinary methodology to help in the design and implementation of computational cognitive models of human social behaviours. These kind of computational models have many applications in areas like robotics, multiagent based simulation, computer games and in general, in those areas where having an artificial entity that shows an accurate human social behaviour is a necessity. The workflow makes special emphasis in the interaction between artificial intelligence, social sciences and humanities as a mechanism to obtain much more realistic social behaviours.

1 Motivation

There are many applications where it is necessary that an artificial entity exhibits a behaviour as close as possible to that of a human, ranging from agent based social simulations or NPCs (non-player characters) in computer games to virtual personal advisors and robot companions just to name a few.

It is clear that in order to build such kind of artificial entities, social sciences and humanities should play an important role. This article presents the workflow of a methodology that can help researchers in computer science, and specifically in artificial intelligence, to build a bridge toward social sciences and humanities as a source for inspiration and validation of computational cognitive models. Although it can be easily understood from the perspective of other disciplines, the methodology is presented having in mind the artificial intelligence researcher and the ultimate goal of embed/integrate computational cognitive models (with a solid theoretical support from social sciences and humanities) in an artificial entity.

We are not claiming that a methodology is the final solution for the problem; there are many other reasons that make interdisciplinary approaches difficult. However we think that a methodology can encourage many researchers to chose this interdisciplinary path and discover that an interdisciplinary approach involving engineering, social sciences and humanities is not an impossible task and that the benefits far exceeds the difficulties. This paper is a first tiny step in that direction.

2 Definitions and observations

- *Methodology.* "A guideline system for solving a problem, with specific components such as phases, tasks, methods, techniques and tools."[2]. As you will notice later, the article you are reading is far from describing a methodology. What is presented is an initial workflow with the main phases for a future methodology. This workflow is intended to act as a roadmap for further development of the methodology based on a set of testing scenarios that are described in section 4.1.
- *Social Cognitive model.* We define a "Social Cognitive Model" as a model (a description of a system, theory, or phenomenon that accounts for its known or inferred properties and may be used for further study of its characteristics[3]) of a mental process that drives a social behaviour.
- *Computational cognitive model.* A cognitive model that presents process details using algorithmic descriptions and can be implemented in a computer.
- *Social behaviour.* Behaviour directed toward society, or taking place between members of the same species.

We distinguish between the different contributions a given discipline can perform in the methodology:

- *Theoretical.* Contributes to the methodology with theories and models of human behaviour. In this area we will usually find disciplines from social sciences (psychology, sociology, anthropology...) and humanities (Philosophy).
- *Engineering.* Any discipline contributing to the actual formalisation and implementation of the models. Examples are Artificial Intelligence, Robotics, Supercomputing or Human computer interaction among others.
- *Empirical.* Disciplines contributing with experimental methods for the validation of the model. For instance Experimental Economics, Experimental Psychology, etc.

For instance, we will identify a given discipline as *theoretical* in a certain stage of the methodology if it is contributing from a theoretical perspective in that stage (that is, providing theories, models, etc.). It has to be clear that this label is only for that stage of the methodology. It doesn't mean that the discipline is theoretical in essence. In fact, it is not strange to find some disciplines that can contribute from several perspectives in different stages. The rationale behind this division should be clear once we present in the next section the workflow of the methodology.

3 The methodology workflow

Figure 1 shows the general schema of the methodology workflow. There are four stages that compound the workflow that will be analysed one by one in the following sub-sections.

[1] IIIA - CSIC, Barcelona-Spain, email: jsabater@iiia.csic.es

[2] http://en.wikipedia.org/

[3] www.thefreedictionary.com

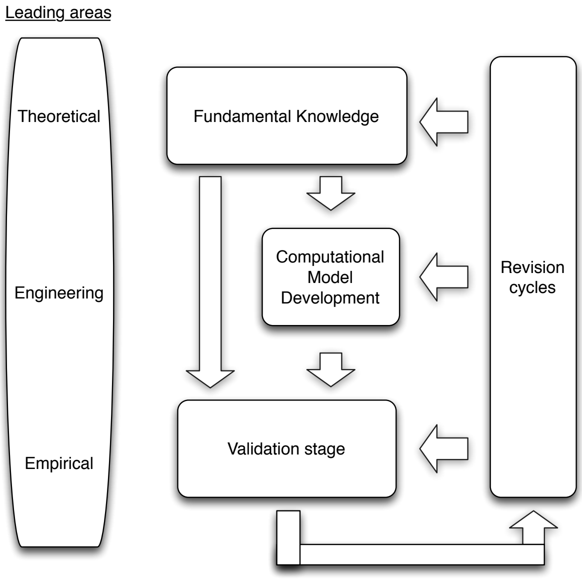


Figure 1. General schema of the methodology workflow.

3.1 Fundamental knowledge

This is the starting point. Once decided the behaviour that needs to be modelled, the first step is to resort to some disciplines that can provide knowledge, theories or ideally cognitive models of that behaviour (see figure 2). At this point it is interesting to consider different points of view coming from several disciplines. Theoretical disciplines should lead this stage.

This step can be trivial if there are cognitive models already available, however the reality usually will be that the available theories and models will be probably too general and vague to move to the next step. Therefore an initial work of refinement involving theoretical and engineering disciplines will be necessary to obtain a cognitive model that can be a good source for the next stages.

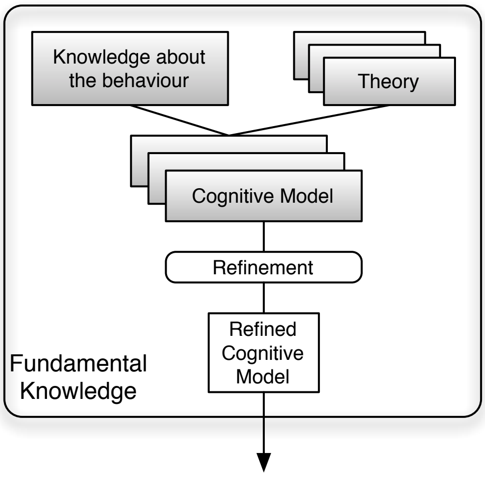


Figure 2. Fundamental knowledge stage.

3.2 Computational models

Once we have a cognitive model detailed enough, we are ready to move to a computational version of it and its integration into a virtual entity. This requires several steps as shown in figure 3). This stage will be lead by the engineering disciplines but always with the support of the theoretical disciplines, specially in the formalisation step. It is not strange that some adjustments to the original cognitive model coming from the Fundamental Knowledge stage need to be done during the process.

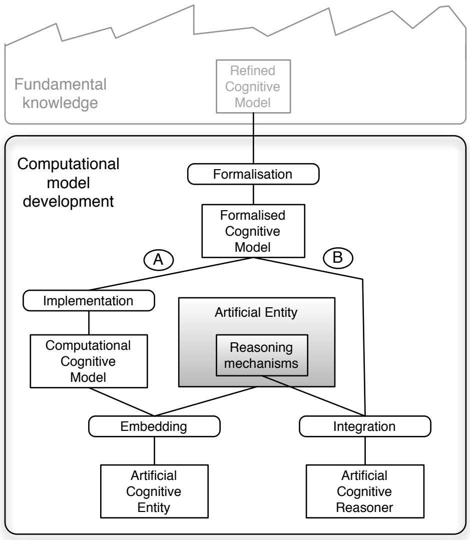

Fundamental knowledge
Refined Cognitive Model
Computational model development
Formalisation
Formalised Cognitive Model
A
B
Implementation
Computational Cognitive Model
Artificial Entity
Reasoning mechanisms
Embedding
Integration
Artificial Cognitive Entity
Artificial Cognitive Reasoner

Figure 3. Computational Model Development stage.

3.2.1 Formalisation

In the previous stage, engineers and theoretical scientists were working together to get a detailed version of the cognitive model. This model, however, is still too rough to be implemented. Although many times (usually due to time constraints) the formalisation step is skipped and engineers go directly to the implementation step, our experience demonstrates that the formalisation exercise is crucial for several reasons:

- It raises to the surface many hidden problems that were not obvious in the textual description of the model.
- It removes any trace of ambiguity.
- It allows to be sure that everybody understand the same about how the model works.

One important aspect to be considered in this step is which formal language to use for the formalisation given the final usage of the cognitive model: (1) If the model will be implemented and embedded into an artificial entity (path A in figure 3) or (2) if the model will be a knowledge source for an automated reasoning process (path B in figure 3). Notice that one usage do not exclude the other.

3.2.2 Implementation, embedding and integration

After the implementation of the formalised cognitive model we will obtain a computational cognitive model ready to be embedded into an artificial entity. Usually it is the computational cognitive model that will have to adapt to the artificial entity (think for instance in a robot platform, a computer game engine with a predefined NPC architecture or a multiagent systems platform with its own agent architecture). Therefore, the knowledge about the final target platform where the computational cognitive model will be embedded should lead the implementation decisions.

It can be the case that we want the cognitive model to be used as part of the knowledge the virtual entity has about the world so it can be incorporated into its reasoning processes. The formalisation of the cognitive model is a first and important step toward the integration of the knowledge present in the cognitive model into the reasoning mechanisms of the artificial entity. The idea is that the artificial entity can use that knowledge also to model the behaviour of other artificial entities and humans. The integration is not concerned only on how to represent the knowledge in a suitable form for the reasoning mechanism but also how to extend the reasoning mechanism to exploit it.

Both, the embedding and the integration require an important bunch of technical work but they also can imply a revision of the cognitive model. Can be that the technical limitations of the artificial entity make impossible the full development of the theoretical cognitive model and therefore a simplification, trying to maintain as much as possible the essence of the original model, is necessary. So although this stage is eminently an engineering task, it can require the participation of the theoretical disciplines to adapt the model. In the worst case, it could happen that the necessary simplifications distort completely the theoretical model till the point it is no longer a valid model.

3.3 Validation stage

The computational cognitive model already embedded/integrated into the artificial entity is based on a more or less solid theoretical background and has been agreed with experts in the theoretical disciplines. However the theories behind the model, the current implementation, the embedding/integration into the artificial entity can be wrong. What we call the *validation stage* is intended to validate the computational model and detect possible problems in the ongoing process.

It is important to notice that our interest is only to check that the behaviour of the artificial entity that is using the computational cognitive model is as similar as possible to the behaviour of a human in the same situation. We will not enter into the controversy if this is enough to validate the initial cognitive model.

There are many possibilities to try to validate the computational model. We propose two of these possibilities based on an empirical approach.

If in the *Fundational Knowledge* stage the theoretical disciplines were leading the process and in the *Computational Models* stage were the engineering disciplines that assumed this role. In the validation stage the lead is for the empirical disciplines.

3.3.1 A scientific control based validation approach

All the process till now was directed to have and artificial entity that can be able to display a specific social behaviour in a similar way a human would do in the same situation. Therefore, the most direct approach to validate the computational model is to use an experimental setting that relies on humans to evaluate if that behaviour is believable or not.

Figure 4 shows this validation stage in a schematic way. We have to design an evaluation scenario where humans can observe both artificial entities with and without the cognitive model as well as other humans impersonating artificial entities. The artificial entities without the cognitive model are the control group, the humans impersonating artificial entities are what we call the reference group and the artificial entities with the cognitive model, the experimental group.

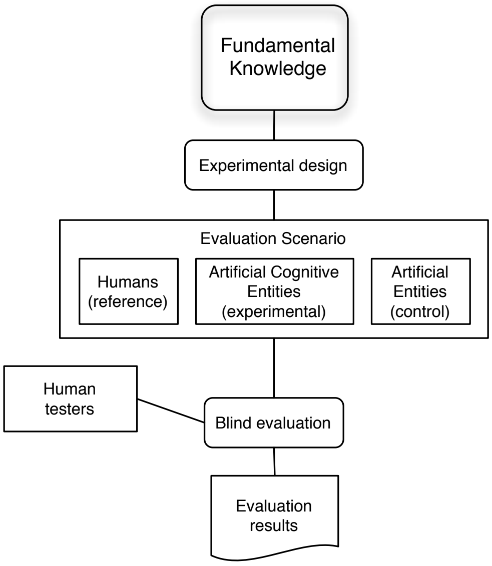


Figure 4. Scientific control based validation approach.

Similar to what happens in any scientific control experiment, using a blind approach (that is, human evaluators do not know with which kind of entity are they interacting with) you should be able to observe how the experimental group and the reference group become indistinguishable and the control group is detected by the human testers.

This approach assumes you can perform the evaluation in an scenario where it is not obvious for the testers to which group (reference, control or experimental) the evaluated entity belongs to. Examples of artificial entities that can be easily evaluated using this approach are NPCs in video games or recommenders and personal advisors.

3.3.2 The comparison between humans and artificial entities

A second approach for the validation is depicted in figure 5. In this approach, the first step is to design an experiment that characterises the social behaviour that has been modelled and that can be run both with humans on one side and artificial entities with the cognitive model on the other. It has to be the same experiment. The idea then is to compare the results of both experiments using the appropriate metrics. If the results are similar, this means that the cognitive model ”behaves” like a human. As you can imagine, the key points in this approach are the design of the experiment and the metrics used to compare the results. Here, the theoretical disciplines play a crucial role, both to design the experiments together with the empirical disciplines and to suggest relevant metrics.

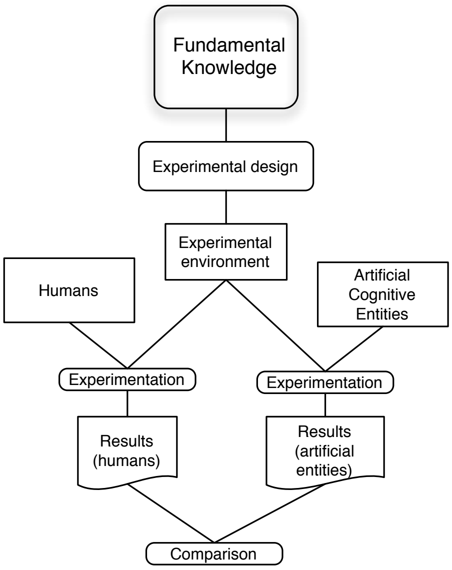


Figure 5. Comparison between humans and artificial entities for the validation.

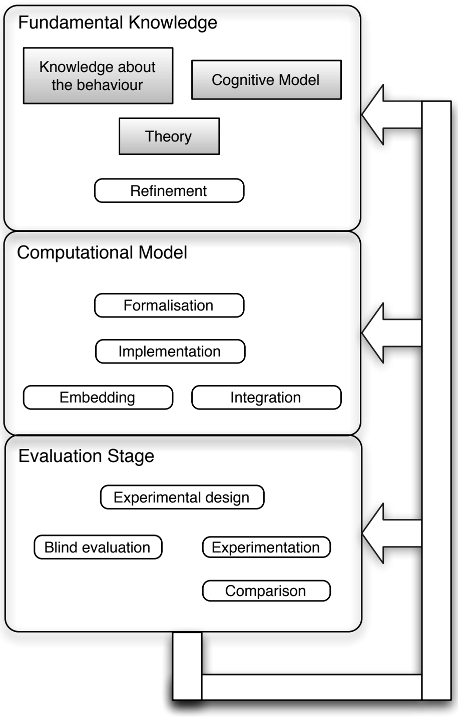


Figure 6. The revision cycles.

3.4 Revision cycles

After the validation stage, it is time to analyse the situation. Probably the validation stage has made explicit that our computational cognitive model is not as human-like as we would like. Figure 6 depicts what we call the *revision cycles*. These are the points where the methodology has to be revised looking for errors. Although every specific context will suggest the best order to follow in order to detect the mistakes, a bottom-up approach (as illustrated in the picture) is usually a good strategy.

4 The methodology to build the methodology

What we have presented till now is only the workflow of a methodology. It has to be considered nothing more than a roadmap. This will be our starting point and the procedure to follow is simple: let's use the methodology as it is in real cases following a "learn by usage" approach. The key point is to select a set of testing scenarios so each one stresses different parts of the methodology, overlapping with the other testing scenarios in some of the other parts. The idea is that each testing scenario can contribute to the refinement of different aspects of the methodology and that the sum of all the testing scenarios allows to cover the whole methodology.

To illustrate this approach we present an example based on four testing scenarios that have been selected following the previous premises.

4.1 Choosing the testing scenarios

Tables 1, 2, 3 and 4 show the summary cards of the testing scenarios. The fields of study try to cover what we think are four main application fields for computational cognitive models: cognitive robotics (testing scenario 1), recommenders and personal advisors (testing scenario 2), multiagent based simulation (testing scenario 3) and computer games (testing scenario 4). The topics of study and specific problems require the participation of a broad range of disciplines. This diversity should give us a good sample of the problems that arise in this kind of interdisciplinary projects and that are associated to the idiosyncrasy of each discipline. It has been taken into account also the kind of interaction that the testing scenario implies, trying to cover as much alternatives as possible (*Interaction* field).

Regarding the coverage of all the steps in the methodology, the alternatives are found in the validation method and the final usage of the computational model. For the validation we have proposed two methods: a *scientific control* method based on reference, control and experimental groups of subjects; and a method that consists on the *comparison* of the results obtained in a set of experiments performed by human beings and the same experiments performed by artificial entities. The four testing scenarios cover both approaches, mixing both methods in two of the cases (see *Validation/Validation method* field in the cards).

Finally, there are two possible usages of the cognitive model contemplated in the methodology: (i) the cognitive model being embedded in the artificial entity regulating its social behaviour (*Embedded* label in the *Comp. Model* field in the cards) and the cognitive model as a source of knowledge for the reasoning mechanism of the artificial entity (*Integrated* label). Again, the different testing scenarios cover the different possibilities, considering in the first testing scenario both alternatives at the same time.

One important aspect that share the four testing scenarios is the necessity of lightweight computational models. In cognitive robotics there is the limitation imposed by the hardware of the robot and the necessity to interact with the real world. Something similar happens in the case of personal advisors and recommenders that should be able to run in mobile devices. In the case of multiagent based simulation, the limits are imposed by the number of artificial entities that need to be simulated in parallel. Finally in computer games it has to be possible to run the models in a single computer, sharing the available computational resources with many other aspects (graphics, general logic of the game, communication, etc.), all of this in real-time.

5 Conclusions and future work

The methodology presented in this paper is still far from being a methodology that can be used out of the box in real cases. Many notions and processes are still too vague and coarse grained and at this moment, are only based on the personal experience of the author after participating in several interdisciplinary projects and interacting with researchers in social sciences and humanities. The intention of this paper was not, therefore, to present a finished methodology but to present the starting point and a roadmap to improve and validate this initial proposal. The next step is to start developing the testing scenarios and learn from the experience in order to refine the proposed stages of the methodology, add/remove stages when necessary and, what is the most important task, listen carefully what researchers from other disciplines have to say.

ACKNOWLEDGEMENTS

This work has been partially supported by the CSIC [Proyecto Intramural de Frontera MacNorms, PIFCOO-08-00017]; the Generalitat de Catalunya [2009-SGR-1434], the European Network for Social Intelligence (SINTELNET), the CBIT project (TIN2010-16306), the AT project (CONSOLIDER CSD2007-0022) and the SmartExFood project (INNPACTO 2012).

Table 1. Testing scenario - 1: Cognitive Robotics

Field	Cognitive Robotics	
Topic of study	Trustworthy robot peers	
Main goal	Design and implementation of a computational cognitive model that allows a robot to display a socially accepted behaviour in front of a mistake.	
Description	How to make technology trustworthy at the eyes of the users has been a matter of study from different disciplines in the last few years. It is not enough that the technology be trustworthy, it has to appear as such at the eyes of the user. This is especially relevant for those technologies that are supposed to behave in a human-like way. We are used to interact with other humans and evaluate the trust we have in them to take decisions. Therefore it is not strange that when we have to interact with artificial entities (an especially if they have a physical body like in the case of robots), we use the same strategies to feed our decision-making mechanism. The problem arises when we are not able to use the same methods to make the trust evaluation of the artificial entity because what we have in front is an entity that behaves in a different way (most probably in a non-readable way). The result is a complete lack of trust in that entity. How to make robots trustworthy is a very broad topic. Human trust is associated to many different behaviours, almost all of them possible in a robot. In this testing scenario we will focus on one of these behaviours: the behaviour in front of a mistake. It has been studied that trust is affected by drops in system's reliability. However, research on ways to recover trust (besides reducing or consistently repeating the failure) is still an open question. How should a robot behave after making a mistake? How can a robot minimise the effect of that mistake from the perspective of trust? Displaying the right behaviour after a mistake can even improve your image in front of a third party, and therefore increase his/her trust on you. On the contrary, the wrong behaviour after even a small mistake can ruin your trustworthiness.	
Interaction	1 artificial entity ⟷ 1 human	
Disciplines	Theoretical	Social Psychology Philosophy
	Engineering	Robotics Artificial Intelligence
	Empirical	Social Psychology
Validation	Validation method	Scientific Control Comparison
Comp. Model	Model usage	Embedded Integrated

Table 2. Testing scenario - 2: Recommenders and personal advisors

Field	Recommenders and personal advisors	
Topic of study	Customer behaviour (information search)	
Main goal	Design and implementation of a computational cognitive model that allows a personal advisor to help the user to find the kind of information he/she is looking for regarding products and experiences associated to those products in a more personalised way.	
Description	The use of artificial entities that act as recommenders and personal advisors in electronic commerce environments is already a reality. One of the main problems that the designers of these systems have to deal with is how to manage the interaction with the user. The interaction with the user cannot longer be a flooding of marketing messages, specially now that with the use of mobile devices, this artificial entities are not restricted to interact with the user only at home but also when the user is on the go at any moment. Fuelled by the ideas of Relationship marketing that recognises the long term value of customer relationship and tries to extend the communication with the customer beyond intrusive advertising and sales promotional, it has appeared the necessity of a new generation of recommenders and personal advisors. These new artificial entities need to have a deep knowledge of the user habits and preferences to ”know” when, how and what kind of information to provide regarding not only new or similar products but also associated experiences. From the six stages of the Consumer Buying Decision Process, this testing scenario will focus therefore on the Information search stage. The idea is that the artificial entity can use different cognitive models of consumer information seeking behaviour. Then applying this knowledge, the artificial entity should be able to fit much better the requirements of information of a particular customer.	
Interaction	1 artificial entity $\longleftrightarrow n$ humans	
Disciplines	Theoretical	Marketing (Consumer behaviour)
	Engineering	Artificial Intelligence Human-computer interaction
	Empirical	Experimental Economics
Validation	Validation method	Scientific Control Comparison
Comp. Model	Model usage	Integrated

Table 3. Testing scenario - 3: Multiagent Based Social Simulation

Field	Multiagent Based Social Simulation	
Topic of study	Observance of social norms in societies without a central authority.	
Main goal	Design and implementation of a computational cognitive model that allows an autonomous agent in a multiagent based simulation to behave like a human regarding the observance of social norms in a society without a central authority.	
Description	The goal of this testing scenario starts from the necessity expressed by researchers in Ethnoarchaeology to use multi-agent based simulation to try to validate different hypothesis about the observance of social norms that govern fisher-hunter-gatherer societies regulating the interactions among the individuals. These societies are characterised for not having centralised institutions that watch over the observance of norms. The specific hypothesis that the simulation intents to validate are out of the scope of this short description. However, a key point in the simulation is to have autonomous agents that can display a behaviour in front of social norms (their observance) as similar as possible to that of humans. It is known that punishment is a key mechanism to achieve the necessary social control and to enforce social norms in a self-regulated society (Axelrod, 1986). On the other hand, the notion of prestige (reputation) has also a great influence on the observance of norms, specially when there is not a central authority that can impose the norms by force. The testing scenario will focus on developing and implementing a computational cognitive model of how punishment and the notion of prestige affect the observance of social norms in societies that do not have a central authority.	
Interaction	n artificial entities $\longleftrightarrow$ n artificial entities	
Disciplines	Theoretical	Anthropology, Ethnoarchaeology Social Psychology Philosophy
	Engineering	Artificial Intelligence Supercomputing
	Empirical	Experimental Economics Social Psychology
Validation	Validation method	Comparison
Comp. Model	Model usage	Embedded

Table 4. Testing scenario - 4: Computer Games

Field	Computer Games	
Topic of study	Believable NPCs (non-player characters)	
Main goal	Design and implementation of a computational cognitive model that, once embedded into a set of NPCs, give them the capacity of spreading rumours (gossiping).	
Description	The goal of this testing scenario is to develop and implement a computational cognitive model of how humans deal with the spreading of rumours. This model, once embedded into a population of NPCs in a virtual environment (for example an MMORPG), should allow the display of this social behaviour in a believable way (from the perspective of the gamers). The NPCs should react to the actions performed by the user and also the actions of other NPCs and spread the rumour as it would happen in a human society.	
Interaction	n artificial entities $\longleftrightarrow$ n humans	
Disciplines	Theoretical	Cognitive sciences
	Engineering	Artificial Intelligence Computer Games programming
	Empirical	Experimental Psychology
Validation	Validation method	Scientific Control
Comp. Model	Model usage	Embedded

A simple logic of tool manipulation
(extended abstract)

Nicolas Troquard[1]

Abstract. Tools are viewed in this extend abstract as artefactual agents: agents whose goals, or function, have been attributed. We put forward an interpretation of tool usage as a social interaction between the tool and its user. Precisely, this social interaction is one of where the tool assists the user to bring about something. We lay out the first principles for a logical approach to reason about the creation and the use of tools. We also discuss some meta-logical properties of the framework.

1 Introduction

Technology is pervasive in our social environment. So much that our societies have been regarded as a huge socio-technical systems. Hence, there is an increasing need for rigorous methods to reason about socio-technical systems, model them, and verify them against a non-ambiguous specification. As formal logics have been successfully applied to the engineering of distributed systems in computer science and electronics, it seems natural to capitalize on them for engineering socio-technical systems as well.

Socio-technical systems are systems where agents in a general sense (entities capable of autonomous choices), interact with designed artefacts. Of these designed artefact, the artefactual agents, or *tools*, are especially relevant to understand the interactions in our societies. The present abstract lays out the first principles for a logical approach to reason about the creation and the use of tools.

The paradigm of multi-agent systems is general enough to encompass socio-technical systems. A tool can be seen as a particular kind of agent: one whose *function*, or goal (or still *telos*, in Aristotle's terminology) has been designed. The function of a tool is to bring about some state of the world when manipulated in a certain manner. Put another way, the function of a tool is to achieve something reactively to the agency of a user agent. We discuss this in Section 3.

Here, our study is formal. We build our logical framework upon Kanger, Pörn, and others' logic of *bringing-it-about*, that we review in Section 2. It already allows to represent in a rigorous manner events of function attribution, and events of actual usage. The full logic extends the logic of bringing-it-about with the means to talk about temporal statements. Prominently, it allows to express the properties that govern the life-cycle of a function of a tool, from its coming into existence to its destruction. We address this in Section 4.

The next section covers the foundations of the logical framework we use to reason about tool manipulation. The reader familiar with the philosophical and formal aspects of logics of agency may only browse it quickly as it contains no original research. A reader unfamiliar even with logical arguments may work the courage and maybe understand, if only a bit, the whys and hows of these specific logics for multi-agent systems.

2 Bringing-it-about logic of agency

Logics of agency are the logics of modalities E_x for where x is an acting entity, and $E_x\phi$ reads "x brings about ϕ", or "x sees to it that ϕ". This tradition in logics of action comes from the observation that action is better explained by what it brings about. It is a particularly adequate view for *ex post acto* reasoning. In a linguistic analysis of action sentences, Belnap and others ([1, 2]) adopt the *paraphrase thesis*: a sentence ϕ is agentive for some acting entity x if it can be rephrased as x sees to it that ϕ. Under this assumption, all actions can be captured with the abstract modality. It is regarded as an umbrella concept for direct or indirect actions, performed to achieve a goal, maintaining one, or refraining from one.

In this paper, we will use the logics of bringing-it-about (BIAT). It has been studied over several decades in philosophy of action, law, and in multi-agent systems ([10], [12], [11], [5], [14], [15], [6], [13], [9], [19]). Following [15], we will then integrate one modality A_x (originally noted H_x) for every acting entity x, and $A_x\phi$ reads "x tries to bring about ϕ".

The philosophy that grounds the logic was carefully discussed by Elgesem in [5]. Suggested to him by Pörn, Elgesem borrows from theoretical neuroscientist Sommerhoff ([16]) the idea that agency is the actual bringing about of a goal towards which an activity is directed. Elgesem's analysis leans also on Frankfurt ([8, Chap. 6]) according to whom, the pertinent aspect of agency is the manifestation of the agent's guidance (or control) towards a goal.

One needs a set of agents Agt and a set of atomic propositions Atm. The language of BIAT extends the language of propositional logic over Atm, with one operator E_i and one operator A_i for every agent $i \in \mathsf{Agt}$. The formula $\phi \wedge \psi$ means that the property ϕ holds and ψ holds. The formula $\neg\phi$ means that the property ϕ does not hold. The remaining logical connectives can be defined in terms of "$\wedge$" and "$\neg$". The formula $\phi \vee \psi$ means that either the property ϕ holds or the property ψ holds. The formula $\phi \to \psi$ means that if it is the case that ϕ then it is also the case that ψ. The formula $\phi \leftrightarrow \psi$ indicates that the previous implication holds and so does $\psi \to \phi$. We use $\top$ to represent a tautological truth.

Formally, the language L is defined by the following grammar:

$$\phi \quad ::= \quad p \quad | \quad \neg\phi \quad | \quad \phi \wedge \phi \quad | \quad E_i\phi \quad | \quad A_i\phi$$

where $p \in \mathsf{Atm}$, and $i \in \mathsf{Agt}$.

A formula of the language is a convenient and rigorous way to characterise properties of interactions between agents. For instance,

[1] LOA-ISTC-CNR Trento, Italy, email: troquard@loa.istc.cnt.it

imagine that deadcoyote represents the property of a world where the coyote is dead. The formula $(E_iA_j\mathsf{deadcoyote}) \wedge \neg\mathsf{deadcoyote}$ then represents the property that agent i brings about that the agent j attempts to brings about that the coyote is dead, and the coyote is not dead.

For any formula ϕ of L, we write $\vdash \phi$ to mean that ϕ is a theorem of the logic. The base principles of BIAT (where i is an individual agent) are:

(prop)	$\vdash \phi$, when ϕ is a classical tautology
(notaut)	$\vdash \neg E_i\top$
(success)	$\vdash E_i\phi \to \phi$
(aggreg)	$\vdash E_i\phi \wedge E_i\psi \to E_i(\phi \wedge \psi)$
(attempt)	$\vdash E_i\phi \to A_i\phi$
(ree)	if $\vdash \phi \leftrightarrow \psi$ then $\vdash E_i\phi \leftrightarrow E_i\psi$
(rea)	if $\vdash \phi \leftrightarrow \psi$ then $\vdash A_i\phi \leftrightarrow A_i\psi$

The set of all the previous principles is the axiomatics of the logic of bringing-it-about. Every base principle captures a key logical aspect of agency. BIAT extends propositional classical logic (prop). An acting entity never exercises control towards a tautology (notaut). Agency is an achievement, that is, the culmination of a successful action (success). Agency aggregates (aggreg). Every actual agency requires an attempt (attempt). The agency (resp. attempt) for a property is equivalent to the agency (resp. attempt) for any equivalent property (ree) (resp. (rea)). So, shaking hand with Zorro is equivalent to shaking hand with Don Diego Vega. Trying to spot the morning star is equivalent to trying to spot the evening star, and it is equivalent to trying to spot Venus.

The decidability of BIAT is important for its practical application in reasoning about socio-technical procedures. The proof is an adaptation of the fact that the satisfiability problem of the minimal modal logic with (aggreg) is PSPACE-complete. (See, e.g., [20].) The full proof for the fragment without the A_i operators is presented in [17]. Completing the proof is straightforward.

Proposition 1 *Let a formula ϕ in the language of BIAT. The problem of deciding whether $\vdash \phi$ is decidable. It is PSPACE-complete.*

This means that we can algorithmically decide of the validity of any property expressed in the language of BIAT. To put it bluntly, a computer can automatically reason for us about properties of action and attempts of agents.

3 Tool function and usage

We may assume that some agents in Agt are acting entities in the general sense, while others are artefactual agents. In the interest of simplicity, in this extended abstract we will assume that we have exactly one particular agent u that we call a "user", and exactly one particular artefactual agent t that we call a "tool".

Tool function. The nature of the activity of a tool is reactive to the (tentative) activity of a user. Hence, the activity of a tool is directed towards goals of the form:

$$A_u\phi \to \phi$$

That is, the *telos* or goal of a tool is "if it is the case that the user attempts ϕ then it is the case that ϕ".

The tool actually exercises its control over such a goal when it brings it about:

$$E_t(A_u\phi \to \phi)$$

Tool usage. We formalise an event of tool usage as an event in which a tool *assists* a user to obtain a goal. The description of the event "the user u achieves ϕ by using the tool t" is as follows.

$$[u : t]\phi \overset{\text{def}}{=} E_t(A_u\phi \to \phi) \wedge A_u\phi$$

So u achieves ϕ by using t when t has the function to bring about ϕ whenever u attempts ϕ, and u attempts ϕ.

This pattern is a particular instance of a more general one. In [3], we use the general pattern to study assistance and help between two acting social entities. In fact, this very pattern is a case of assistance.

It is a *successful* use because we have the following expected property by applying (success) and (prop):

Proposition 2 $\vdash [u : t]\phi \to \phi$

It is an assistance event for three reasons. First, there is an *assistee*, the user. It is a goal of u to bring about ϕ and u does try. Second, there is an *assistant*, the tool. t's guidance is reactive to u's goodwill in the action. Here, the goal of u is that ϕ holds if j tries to bring about ϕ. Third, despite Prop. 2, it is the case that $[u : t]\phi \wedge \neg E_u\phi \wedge \neg E_t\phi$ is a consistent formula. That is, it is possible that t successfully assists u to bring about ϕ, and still, neither t nor u brings about ϕ. Hence, the success of the event of tool usage described by $[t : u]\phi$ comes from some cohesion between u and t. (This cohesion is exploited in [18] to characterise group agency in BIAT.)

Grounding the user's attempts. It might seem rather arbitrary to reduce the usage of a tool to achieve ϕ, to u's attempt to achieve ϕ. This is a harmless simplification which abstracts away from the actual manipulations of the tool that the user must perform to use its functions. For instance, if t is a gun, the user might need to pull the trigger for the gun to fire and kill the coyote: this would correspond to the function $E_t(E_u\mathsf{trigger} \to \mathsf{deadcoyote})$.

Now, the fact that the user kills the coyote by using the gun is captured by:

$$E_t(E_u\mathsf{trigger} \to \mathsf{deadcoyote}) \wedge E_u\mathsf{trigger}$$

The gap between the specific manipulation of the gun and the attempt to kill the coyote can be filled in the logical theory. For instance, by stipulating the following:

$$A_u\mathsf{deadcoyote} \leftrightarrow E_u\mathsf{trigger} \vee E_u\mathsf{rope} \vee \ldots$$

It explains u's attempt of killing the coyote as the act of pulling the trigger, or passing a rope around the coyote's neck (rope), or possibly doing other relevant actions.

4 Tools as agents with designed functions

A tool is an artefact. It is what it does, and it does so because its function has been designed and attributed by a creator. In our simple setting, the user will also be the creator.

To express the properties pertaining to the existence of a tool function and the persistence of a tool function we will use the additional expressiveness of tense logics. In the following $\phi\mathcal{S}\psi$ reads that ϕ holds ever since ϕ does; $\phi\mathcal{U}^w\psi$ reads that ϕ holds until ϕ does, or ψ never occurs. ($\mathcal{U}^w$ is the *weak* until of tense logic.) At the end of this section we briefly discuss the technicalities concerning the addition of the temporal dimension.

Attributing a function. The logic can express that u attributes the function of assisting her to achieve ϕ as follows:

$$E_u E_t(A_u\phi \to \phi)$$

So, u brings about that t brings about that ϕ holds whenever u tries to achieve ϕ.

Existence of a function. A tool is an artefact. Its functions have been designed by the creator/user. We adopt the following principle.

$$E_t(A_u\phi \to \phi) \to$$
$$(E_t(A_u\phi \to \phi)\mathcal{S}E_u E_t(A_u\phi \to \phi)) \vee (E_u E_t(A_u\phi \to \phi)) \quad (1)$$

In English, if t has a function then either (i) there is a time strictly in the past where u attributed this function to t, and t has consistently held the function ever since, or (ii) u attributes this function to t at the present time.[2]

Persistence of a function. The sort of agency $E_t(A_u\phi \to \phi)$ that a tool has, is different from the sort of agency $E_a\gamma$ that a natural agent a has. If $E_a\gamma$ holds at some time, it is no assurance that $E_a\gamma$ will hold after. The agent a's goals are ever changing and so is her activity towards them. This is different for $E_t(A_u\phi \to \phi)$ because it is intended to reflect some designed function attributed to an artefact.

The activity of a tool persists. At least it persists until its function is altered by u. When a chimp takes out the leaves of a thin branch to use it as a stick and collect ants, the function of the stick will be the same the next hour, and the hour after that. Unless eventually the chimp crushes it. We then adopt the next principle:[3]

$$E_t(A_u\phi \to \phi) \to$$
$$E_t(A_u\phi \to \phi)\mathcal{U}^w E_u\neg(E_t(A_u\phi \to \phi)) \quad (2)$$

Meta-logical analysis. Adding a temporal dimension, we have considerably complicated the logical framework. However, it is in fact easy to provide a rigorous semantics to the new language by using Finger and Gabbay's *temporalisations* ([7]). We can restrict the class of all model to the constraints for which Principle 1 and Principle 2 are canonical, and we obtain the *class of models for tool manipulation*.

Combining the axiomatics of BIAT, the axiomatics of Since-Until tense logic ([4, 21]), Principle 1, and Principle 2, we immediately obtain an axiomatic theory that is sound and complete wrt. the class of models for tool manipulation.

Since BIAT is decidable (Prop 1), and so is Since-Until tense logic, a general result of Finger and Gabbay can even be applied to assert that the reasoning problem in the resulting theory is decidable.

ACKNOWLEDGEMENTS

This research is part of the project *Logical Analysis of Socio-Technical Systems* (LAST) funded by the program Marie Curie People Action COFUND Trentino.

[2] This principle must be adapted accordingly if we have more than one creator in the system.

[3] Again, this principle must be adapted accordingly if we have more than one agent in the system who can alter the function of the tool.

REFERENCES

[1] Nuel Belnap and Michael Perloff, 'Seeing to it that: a canonical form for agentives', *Theoria*, **54**(3), 175–199, (1988).

[2] Nuel Belnap, Michael Perloff, and Ming Xu, *Facing the Future (Agents and Choices in Our Indeterminist World)*, Oxford University Press, 2001.

[3] E. Bottazzi and N. Troquard, 'A philosphical and logical analysis of help'. Working title, 2013.

[4] John P. Burgess, 'Axioms for tense logic. I. "since" and "until"', *Notre Dame J. Formal Logic*, **23**(4), 367–374, (1982).

[5] Dag Elgesem, *Action theory and modal logic*, Ph.D. dissertation, Universitetet i Oslo, 1993.

[6] Dag Elgesem, 'The modal logic of agency', *Nordic J. Philos. Logic*, **2**(2), (1997).

[7] Marcelo Finger and Dov M. Gabbay, 'Adding a temporal dimension to a logic system', *Journal of Logic, Language and Information*, **1**, 203–233, (1992).

[8] Harry Frankfurt, *The Importance of what We Care About*, Cambridge University Press, 1988.

[9] Jonathan Gelati, Antonino Rotolo, Giovanni Sartor, and Guido Governatori, 'Normative autonomy and normative co-ordination: Declarative power, representation, and mandate', *Artificial Intelligence and Law*, **12**, 53–81, (2004).

[10] Stig Kanger and Helle Kanger, 'Rights and Parliamentarism', *Theoria*, **32**, 85–115, (1966).

[11] Lars Lindahl, *Position and Change – A Study in Law and Logic*, D. Reidel, 1977.

[12] Ingmar Pörn, *Action Theory and Social Science: Some Formal Models*, Synthese Library 120, D. Reidel, Dordrecht, 1977.

[13] Lambèr Royakkers, 'Combining deontic and action logics for collective agency', in *Legal Knowledge and Information Systems. Jurix 2000: The Thirteenth Annual Conference*, eds., Joost Breuker, Ronald Leenes, and Radboud Winkels, pp. 135–146. IOS Press, (2000).

[14] Felipe Santos and José Carmo, 'Indirect Action, Influence and Responsibility', in *Proc. of DEON'96*, pp. 194–215. Springer-Verlag, (1996).

[15] Felipe Santos, Andrew Jones, and José Carmo, 'Responsibility for Action in Organisations: a Formal Model', in *Contemporary Action Theory*, eds., G. Holmström-Hintikka and R. Tuomela, volume 1, 333–348, Kluwer, (1997).

[16] Gerd Sommerhoff, 'The Abstract Characteristics of Living Systems', in *Systems Thinking: Selected Readings*, ed., F. E. Emery, Penguin, Harmonsworth, (1969).

[17] N. Troquard, 'Reasoning about coalitional agency and ability in the logics of "bringing-it-about"'. Under review, 2012.

[18] N. Troquard, 'The social fabric of cohesive group agency: an abstract logical framework'. Under review, 2013.

[19] Nicolas Troquard, 'Coalitional Agency and Evidence-based Ability', in *Proc. of AAMAS'12*, eds., Conitzer, Winikoff, Padgham, and van der Hoek, pp. 1245–1246. IFAAMAS, (2012).

[20] Moshe Vardi, 'On the Complexity of Epistemic Reasoning', in *Proc. of Fourth Annual Symposium on Logic in Computer Science (LICS'89)*, pp. 243–252. IEEE Computer Society, (1989).

[21] Ming Xu, 'On some u, s-tense logics', *Journal of Philosophical Logic*, **17**(2), 181–202, (1988).

Towards a Design Framework for Controlled Hybrid Social Games

Harko Verhagen [1] and **Pablo Noriega** [2] and **Mark d'Inverno** [3]

We propose a framework for designing and deploying games where social behaviour is kept under control. This framework may also be used for designing other dynamic coordinated social spaces.

1 INTRODUCTION

In digital game research it has been noted that even though game developers use AI as a selling argument, the intelligence is rather shallow. A close analysis of current game agents (or *non-player characters*, NPCs) reveals many deviations of intelligent behaviour.

According to Bartle [4], NPCs may have several functions in a game:

- buy, sell, and make stuff
- provide services
- guard places
- get killed for loot
- dispense quests (or clues for other NPCs quests)
- supply background information (history, lore, cultural attitudes)
- do stuff for players
- make the place look busy.

Many times, NPCs are unaware of their environment, causing them to miss essential information and die as a result. They are also usually unable to dynamically build temporary coalitions, reason about norms, or use any other social coordination mechanism that will control their decision making and behaviour. In fact, most NPCs are very simple script machines that have very poor social skills. Dynamic NPCs would be one solution to the boredom of having to deal with these simple messengers. However, game designers fear that the loss of control over the agents may cause the game to get out of hand and destroy the gaming experience by breaking the storyline.

Our proposal is that in addition to using different NPC architectures based on the function of NPCs (thus making them more challenging or engaging), one may rely on *social coordination mechanisms* that apply within such hybrid social games (hybrid here implies a mix of human and NPC participants interacting) and thus achieve an acceptable level of control over characters. We understand that both elements are non-trivial and carry different concerns, hence we propose to separate those concerns by proposing a clear separation between the game itself and the characters that participate in it. In this paper we focus on the first direction.

Although in this paper we limit our discussion to hybrid social games, the design approach that we propose applies to other sorts of dynamic coordinated sociotechnical systems like participatory simulation environments and open regulated MAS.

2 RELATED WORK

Some related works exists in both game research (concerning the design framework as such) and participatory simulation research (dealing with hybrid social spaces). In game research, the MDA framework presented in [11] is of interest. The framework is meant to make iterative design processes involving developers, researchers, and designers more easy by distinguishing between the game development from a designer perspective and from a player perspective. In the view of [11], the developer would focus on the game mechanics (M), while the player is more focussed on (or expressed in ideas of) the game aesthetics (E), expressed in the emotions the game produces in the player while playing. The runtime behaviour of the game is called dynamics (D). In contrast to our framework presented below, the MDA framework is including the basic game idea (aesthetic) while remaining under-specified with respect to the mechanics and elements involved and also lacks a separation between the control of the game elements from the design and play perspective. In participatory simulation (such as [1]), the mix of human and nonhuman agents is needed to build a better understanding (or increase knowledge of) a real world situation or system by the human agents. Thus, there is a target system that is known and is characterised by (verified) empirical data and hypotheses. In computer games, the target system does not exist, it is a designed world. Thus the relationships between the game elements cannot be empirically verified nor theoretically grounded in a decisive way.

3 DESIGN FRAMEWORK

3.1 Games and Coordinated Social Spaces

We can see games—and several other coordinated social spaces—as having three complementary and interrelated views.

1. The first view, (circle "I" in Fig. 1) is the *ideal* game where some (ideal) characters interact according to the rules of the game. The"'I" view contains an idealised description of the landscape where the game takes place, refers to knights and aliens or football players that will be involved in the game and exchange messages or kill each other according to the scripts of "Assassin's Creed" or whatever.
2. Another view (square "T"), consist of the *technological artefacts* that implement and support that ideal game. This support is of two sorts: On one hand, there are the technological artefacts that implement and run the ideal game, namely, it includes the code of avatars that will be used by humans to play, as well as the code of NPCs,it makes operational the actions of characters on the game "landscape", and, in general, makes sure that the game is executable and follows the conventions that govern the ideal game.

[1] Stockholm University. Sweden; email:verhagen@dsv.su.se
[2] IIIA-CIC, Spain; email:pablo@iiia.csic.es
[3] Goldsmiths, University of London, UK; email:dinverno@gold.ac.uk

On the other hand, there is the software and hardware needed to support the running game: communications, data-bases, interfaces, etc.

3. Finally, the third view (triangle "W"), is the physical world where the game takes place (the room, the screen, the console and the humans that play the game).

In this paper we will look mainly into what the contents of "I" and "T" should be, and how these two views of a game are related.

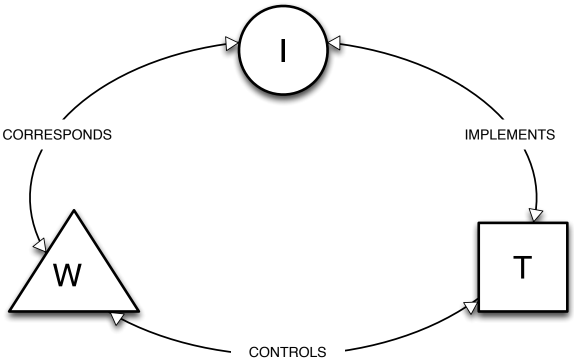


Figure 1. The three views of a coordinated social space: The ideal game, the technological artefacts that implement it and the actual world where the game is played

We propose that the design of a game should take into consideration a number of aspects that become instantiated and assembled to become what we call a *game space*, which is part of the "I" view. The outcome of that assembly will be a precise description of an ideal game that may then be specified, implemented and run within its *game environment* as a technological artefact. The instantiation and assembly of the game components, however, may take different forms if the repertoire of aspects is rich enough. Thus, we claim it is worth separating those elements that may help in choosing the better components from those needed for assembling those components. For that purpose we propose to have a *design space* (again in"I") and couple it with a *design environment* that includes the game environment and the repositories and services that support and complement it (in "T").

3.2 The Game Space

Building loosely on a theatrical metaphor, we may visualise the game space as a play. The author of the play creates a plot and organises it into scenes where some characters exchange dialogue and gestures and move around the staged rooms according to some directions. Likewise, the game space is an abstract entity, designed by someone, where humans and NPCs interact, by means of some interaction mechanisms, within collective activities that happen in particular "rooms" subject to some procedural and behavioural conventions.

More precisely, a game space defines an interaction framework where agents—that may be humans or software and may be created by the game designer with a specific purpose in mind or may be bonafide players—are able to perform certain (admissible) actions subject to some ways of imbuing "acceptable" social order. We propose that in order to define that interaction framework, the designer needs to address ten aspects for which a collection of conceptual constructs are available. In fact, the instantiation of particular constructs and their assembly will constitute the actual game space.

These are the aspects that we believe are necessary and sufficient to describe a game space and we exemplify the type of constructs each should include:

Ontology. It is worth distinguishing between game-generic and game-specific ontologies. In both cases we mean ontology as "entities", or a collection of "terms" in"I" that are eventually mapped into "W".

- *game-generic constructs* are needed to define contexts of collective interaction and their interrelations. For example: action, agent, role, and notably *collective contexts* (ideal locations or activities where several agents interact simultaneously, sharing the *same state*) like game level, scene, transition, challenge, ...
- *game-speciic*. This will list the elements the are used to define the content of collective contexts and "interactions" (actions). For example, swords, ditch, wall; dig, climb, exit, acquire role, improve prestige,...; raise hand, ...

Agent types These include the two main types of "embodied" participants: PCs and NPCs, and perhaps some server agents, which are not visible to players, that deal with some game management functions (for instance performing police-like and time-keeping functions).
Notice that although we want agents to participate in the game, we do not include them as part of the game space. However, we specifically want to distinguish between playing characters and NPC. The former are assumed to be independent of the designer while in the second type, the designer has control over their definition within the *design space* as we shall see below.

Social constructs. Describe the way individuals are related among themselves and also serve as means to refer to individuals and collectives by the role they play rather than by who they actually are. These may include: roles; relations among roles (n-ary relationships between individuals as well as higher-order relationships. i.e, groups, hierarchies of roles, power relationships and so on); organisations (groups plus coordination conventions)

Actions . It is worth distinguishing at least three types: individual actions (pick up a sword, climb a wall, move towards an object); interactions (actions involving two or more agents like attack an enemy, ask for directions, proclaim an outcome) and actions towards game-generic constructs (change a level, embark in a challenge)

Languages. These are needed to define the behaviour of the system and the way it is regulated. These may be organised as a hierarchy of languages that starts with a *domain* language (to refer to the basic game objects: mountains, walls, sword, attire, coins,...) that includes terms of higher *action* languages (description of an action); followed by *constraint* languages (preconditions and post-conditions of actions); then *normative* languages (procedural, functional or operational directions; behavioural rules,...) and so on, depending on the complexity of the definition of the gae and the particular choice of aspects.

Social order constructs. To allow top-down or bottom-up articulation of interactions, the usual device is to use different types of norms: procedural, constitutional, rules of behaviour,...

Social order mechanisms. To allow top-down or bottom-up governance. Among these: regimentation (rendering some actions impossible, strict application of sanctions,...); social devices (trust, reputation, prestige, status, gossip); policing devices (law enforcement),...

Evolution. The game may evolve over time as a result of emerging social conventions, adaptation to different populations of players or to some performance criteria like the quality of engagement or the success rate in some challenges. The definition of the game should include the devices through which that change happens: performance indicators, normative transition functions and such.

Inference. In case the description of situations is somewhat normative, the designer may want to postulate different ways of inferring intended or observed behaviour. For example, *classical logical inference* to allow norm-aware agents to decide whether or not to comply with a norm at some point, to allow police-like NPC to infer a potential misconduct, and so on; reasoning under *uncertainty*; *coherence* as alternative to classical forms of inference when validating game conventions off-line or monitor on-line evolution of a game.

Information structures. that are associated with the main entities of the game, agent profiles and the profiles of active game-specific constructs. In particular, every game needs to keep that information that may change as the game is played: the (shared) state of the system (the value of each and every variable that may change through the action of some agent or the passing of time),

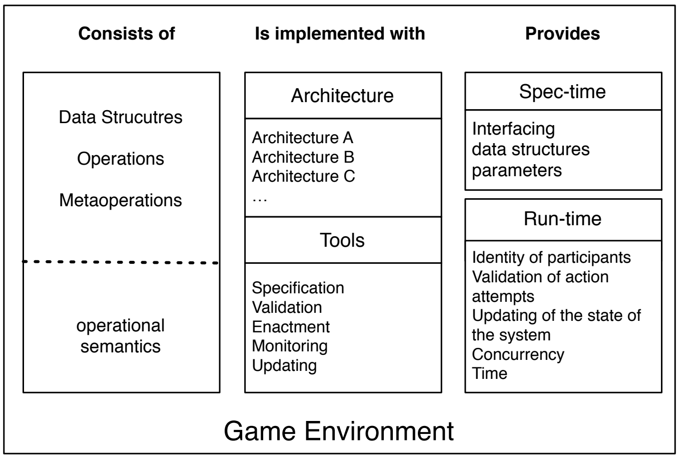


Figure 2. Game Environment

3.3 The Game Environment

While the game space is an ideal description of the game, there should be some device to turn that description into code that allows humans and software agents to be part of an enactment of a hybrid social game. The game environment is made by those technological artefacts that allow that to happen.

As Fig.2 suggests, the game environment contains all the data structures and operations that allow the implementation of the instantiations of all those constructs needed to make a precise description of a game space and the op erational semantics that determine when an input to the system is admitted and its effect.. Thus one needs to have data structures and algorithms to refer to swords and agent types, to denote the attempt to climb a wall or attack and alien ship, to represent challenges and other collective contexts, to establish regimented regulations, allow adequate transparency for social order constructs to apply, and in essence to represent, control, and update the state of the game at every instant the game is being played.

Ideally, there should be a formal definition of these data structures, operations and semantics so that (i) they may be implemented in one or various architectures (centralised, distributed or mixed), (ii) an appropriate specification language may be built and (iii) the corresponding middleware produce a run-time version of the game (in "T") that allow players (in"W") to engage in the game.

3.4 Design Space and Design Environment

While the game space is the ideal game, the designer has to make choices as to what are going to be the actual components of the game in function of some design criteria and assumptions regarding the decision-making capabilities of participants (PCs as well as NPCs), the intended level of complexity, the expected or desired evolution of the game, the purported level of engagement , etc. While we presume that most challenges are dealt with thanks to the repertoire of constructs available for the game space, the design space will contain other constructs to complement the activation of the game space.

All those components, in turn, are then supported by technological artefacts that will eventually constitute the environment where the game is designed, assembled , enacted and maintained. Schematically, as depicted in Fig. 3, the design environment will include:

- The design and activation of NPC and support agents:
- Tools for monitoring and updating the game.
- In general, services and repositories of data and of knowledge. Services like a model to calibrate parts of the game, trust and quality assessment, and in social simulation systems, econometric or demographic models to supplement agent-based models, scenario specification, performance indicators, parameter changing functions,... . Repositories like heuristics for experimental design or calibration of performance indicators, environmental or economic data and so on.
- Finally, one may want to consider as part of the design space other support and management technologies that, being associated with the game, are arguably not part of it. For instance, a forum-like facility to exchange messages among players; a polling device to get opinions about possible game extensions or other games; and of course all the back-office support.

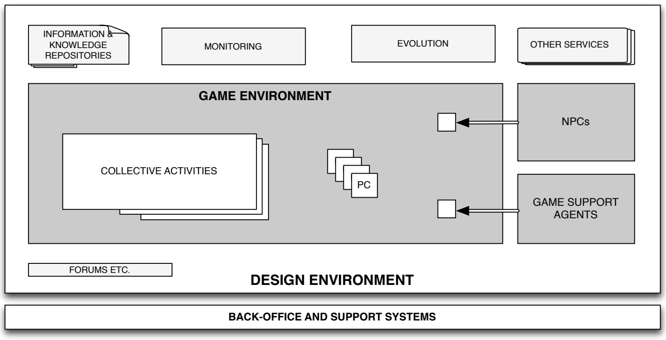


Figure 3. The Design Environment includes the game environment plus those services that make the game playable

4 CLOSING REMARKS

4.1 Separation of concerns

While we motivated this paper as a way of making the use of artificial intelligence more valuable for games, we postulated to address this problem by dividing it in two. We proposed an obvious separation of concerns between design and implementation of those NPC that that the designer wants to have in the game, on one side, and on the other, those of the game milieu and its components. The paper dealt only with the second aspect by making three proposals:

- We propose a three-fold understanding of games (depicted in Fig. 1) and from that picture we proposed a second and perhaps not so obvious, separation of concerns between *game design* and *game implementation*.
- We propose a list of *aspects* that need to be taken into account in the design of a game, propose that for each of these aspects, there may be different ”conceptual constructs” that may be assembled to make those aspects concrete for a given game, and for each of these conceptual constructs and their assembly there should be a *computational counterpart* that implements them.
- Finally we propose that the design and enactment of the actual game requires another layer composed by a design space that includes conceptual means to choose the constructs of the design space and an environment where the implementation of the game is complemented with other artifacts that allow its enactment

We are confident that this approach is one reasonable way of addressing the clumsiness of NPCs without loosing control over the game that we mentioned at the top of the paper. Moreover, it is our impression that game developers take a shortcut from the game or simulation idea to the tools and architectures, then iterate back to the game space. This limits of course the game space contents and constructs used. Computer scientists may use all elements but lack the input from the social science on how to fill the game or deign space with values based on sound theories and empirical work from the social sciences. Thus the contribution potential of both communities to each other is clear. Analysing existing games using these constructs as well as rebuilding them using the design proposal is a natural next step.

4.2 Backing

A substantial influence is the work on electronic institutions. The EI framework (see [8]) is a particular, restricted, version of the framework we propose here. Figure 4 suggest how the constructs we propose for the game space extend the EI ones. Moreover, some of the ideas of the design space are already present in an extension of the EI framework with services for simulation [3]. Likewise, the experience of an implementation architecture and the corresponding EIDE development tools [10], their light-weight variants suggested in [9, 12] and their 3–D extensions [13] give an inkling of what is involved in the tasks ahead.

4.3 Games as a special case of coordinated social spaces

Our game space is but a particular collection of aspects, most of which are also relevant for several types of controlled spaces where many individuals interact in some endeavour that they cannot achieve in isolation. Similarly, the game environment would also have a counterpart in these coordinated social spaces. What may not be so obvious, though is that an analogue of the design space and the design environment are also applicable. An immediate example is the case of participatory agent-based simulation, where a straightforward specialisation of the game space would be the actual simulation model and the environment, the implemented model.

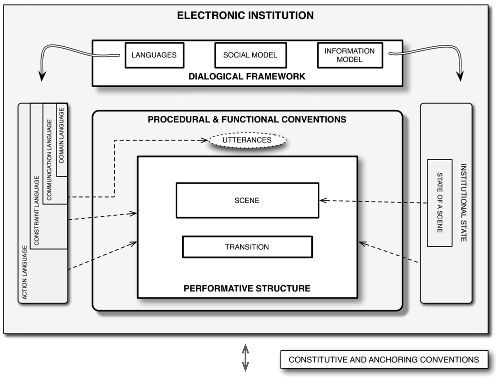


Figure 4. The conceptual space of electronic institutions

One layer up, the design space and environment would include the conceptual tools—statistical techniques, scenario definition, macro-view models—and their technological counterparts—services and repositories—that allow for the design, enactment and analysis of experiments. It would be appropriate to contrast the proposed game space components with the conceptual needs of sociological theories and enrich the game space accordingly.

We have drawn inspiration from the work in normative multiagent systems ([7, 6, 5]. In particular, the reader will notice a closer affinity of this proposal with the proceedings of the Dagstuhl Normas2012 workshop [2] and a version of Fig. 1 and something akin to the design space are included in the forthcoming follow-up volume of the workshop. We claim that our framework would fit nicely with a large number of open regulated multiagent systems.

Finally, we would like to note that the intuitions behind Fig.1 apply not only to hybrid social games or normative MAS but we believe the three-fold correspondence applies to a large variety of sociotechnical systems and our framework could be tuned to the peculiarities of several of them. We propose to abstract from the game space to a space where *agreements* take place and reify the environment as an open environment where computation by agreement is feasible. Perhaps this is one way to move towards the understanding of social intelligence.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments which helped improve this paper. This research was partly funded by SINTEL-NET FET Open Coordinated Action (FP7-ICT-2009-C Project No. 286370) and Consolider AT project CSD2007-0022 INGENIO 2010 of the Spanish Ministry of Science and Innovation.

REFERENCES

[1] Diana F. Adamatti, Jaime Simão Sichman, and Helder Coelho, ‘An analysis of the insertion of virtual players in gmabs methodology using the vip-jogoman prototype’, *Journal of Artificial Societies and Social Simulation*, **12**, (2009).

[2] Giulia Andrighetto, Guido Governatori, Pablo Noriega, and Leon van der Torre, ‘Normative Multi-Agent Systems (Dagstuhl Seminar 12111)’, *Dagstuhl Reports*, **2**(3), 23–49, (2012).

[3] Josep Lluis Arcos, Pablo Noriega, Juan A. Rodriguez-Aguilar, and Carles Sierra, ‘E4mas through electronic institutions.’, in *Environments for Multi-Agent Systems III.*, eds., D. Weyns, H.V.D. Parunak, and F. Michel, number 4389 in Lecture Notes in Computer Science, 184–202, Springer, Berlin / Heidelberg, (08/05/2006 2007).

[4] Richard A. Bartle, *Designing virtual worlds*, New Riders., 2003.

[5] Guido Boella, Pablo Noriega, Gabriella Pigozzi, and Harko Verhagen, eds. *Normative Multi-Agent Systems*, volume 09121 of *Dagstuhl Seminar Proceedings*, Dagstuhl, Germany, 15/03/2009 2009. Schloss Dagstuhl, Leibniz-Zentrum fuer Informatik, Germany.

[6] Guido Boella, Gabriella Pigozzi, and Leender Van der Torre, 'Five guidelines for normative multiagent systems', in *Legal Knowledge and Information Systems. JURIX 2009*, ed., Guido Governatore, pp. 21–30, Amsterdam, (October 22-24 2009). IOS Press.

[7] Guido Boella, Leendert van der Torre, and Harko Verhagen, 'Introduction to the special issue on normative multiagent systems', *Autonomous Agents and Multi-Agent Systems*, **17**, 1–10, (2008).

[8] Mark d'Inverno, Michael Luck, Pablo Noriega, Juan A. Rodriguez-Aguilar, and Carles Sierra, 'Communicating open systems', *Artificial Intelligence*, **186**(0), 38 – 94, (2012).

[9] Marc Esteva, Juan A. Rodriguez-Aguilar, Josep Lluis Arcos, and Carles Sierra, 'Socially-aware lightweight coordination infrastructures', in *AAMAS'11 12th International Workshop on Agent-Oriented Software Engineering*, pp. 117–128, (2011).

[10] Marc Esteva, Juan A. Rodrguez-Aguilar, Josep Lluis Arcos, Carles Sierra, Pablo Noriega, and Bruno Rosell, 'Electronic Institutions Development Environment', in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '08)*, pp. 1657–1658, Estoril, Portugal, (12/05/2008 2008). International Foundation for Autonomous Agents and Multiagent Systems, ACM Press.

[11] R. Zubek R. Hunicke, M. LeBlanc, 'Mda: A formal approach to game design and game research', in *Proceedings of the Challenges in Game AI Workshop, Nineteenth National Conference on Artificial Intelligence.*, (2004).

[12] David Robertson, 'A lightweight coordination calculus for agent systems', in *Declarative Agent Languages and Technologies. DALT 2004*, volume 3476, pp. 183–197. Springer, (2005).

[13] Tomas Trescak, Inmaculada Rodriguez, Maite López-Sánchez, and Pablo Almajano, 'Execution infrastructure for normative virtual environments', *Engineering applications of artificial intelligence*, **26**, 51–62, (01/2013 2013).

Ethnoarchaeology meets Artificial Intelligence An ongoing experience

Asumpcio Vila, Manuela Perez [1]
Jordi Estevez, [2]
Jordi Sabater-Mir, David de la Cruz, Daniel Villatoro [3]
Adria Vila [4]

Abstract. We describe the problems and issues that have emerged during the development of an application using multiagent systems to test a hypothesis about the development and impact of social norms that affect the reproduction of small scale hunter-gatherer societies.

1 The problem

The current normal archaeological record (prehistoric) does not give direct answers to questions relative to the social organisation (that is, to the kind of relations between people for production and biological and social reproduction). Thus for instance we ignore actually how was the sexual division of labour or if structural disymmetry between sexes did exist in the oldest Prehistory. No hypothesis has been yet verified about the origin of this division and disymmetry and the causes of their continuity or change. These are essential questions if we want to do an objective analysis of the present state, a diagnosis, and a prospective of future possibilities of the social relationships.

1.1 Starting point

From a systematic and objective analysis of the written and graphic sources about small scale ethnographic societies (societies with a subsistence system similar to that of the first fully human societies) we state two universals: sexual division of the activities and structural asymmetry between women and men [4].

These small-scale societies are self-regulated and *without central institutions*[5] *but with strict norms that regulate social behaviour*. Their norms regulate the relations and conform the morphology of the societies. They are not subject to specific institutions but seem to be part of the cultural baggage, of the tradition, of what has always been [3].

1.2 Hypothesis

Structural inequality between men and women is linked to the need of limiting the demographic grow in hunter-gatherer societies. The high reproduction capacity in humans had to be limited. Reducing the number of females that arrive to the reproduction was the most effective way to accomplish that goal. Restrictive social rules, as well as the undervaluation of women were a simple way to arrive to the desired effect. This damage to part of the society (women) in the short term could be supported because the collective advantages at a long term. This is a structural feature in hunter-gatherer societies, because without asymmetry the restriction in reproduction would not have had take place, due to the little control over their own biology.

1.3 Objectives

Objective 1: To establish the basis of a general model of regulated normative social behaviour in small-scale societies without political institutions but with strict social rules.

Specific objective: To study mechanisms of self-organisation and distributed social control (such as reputation / prestige), which are generators, and maintainers of social norms in human societies. This can be translated to virtual environments. We want to study the mechanisms that make a set of rules become dominant and are used (and maintained) by the majority of members of a society. We believe that the existence of such shared norms is the kernel that constitutes the society" and are therefore internalised. We refer to social norms that emerge from a decentralised interaction between members of a collective and are not imposed by any authority [1].

The objective of the simulations is to answer questions like:

- How does the normative system determine the viability of a society?
- What norms are essential to the sustainability of a social system in a specific environment?
- Could other regulatory systems have the same effect in the society in that environment?

Finally

- Contrast the efficiency of norms for the group survival.
- Analyse in which environments are these regulatory mechanisms successful.
- Study the positive and negative consequences on the group.

Objective 2: Refine hypotheses about the mechanisms of assumption, transmission and maintenance of norms that allowed inequality in hunter-gatherer societies. Detect what indicators (the materiality) will enable us to demonstrate the existence that can generate proposals of how to get a proper archaeological record in the study for the origins of the variety in human social relations.

Objective 3: From an artificial intelligence perspective, to develop a simulation platform where the normative system is both the kernel

[1] IMF - CSIC, Spain
[2] UAB, Spain
[3] IIIA - CSIC, Spain
[4] UOC - Girona, Spain
[5] Social organisations devoted to maintain a social order.

and the main research subject of the simulation. In other words, the purpose of the simulations is to answer questions like: How the normative system determines the viability of a society? Which norms are essential for its sustainability in that specific environment? Could other normative systems have the same effect on that society in that environment? How much does the normative system contribute to the sustainability and prosperity of a society? [5, 2]

2 Methodology and challenges

The emergence and operation of such norms is a critical issue that has stimulated repeated trials. But never before the experimentation in Social Sciences has explored the emerging scientific space generated on the confluence of different disciplines like: Ethnography, Archaeology and Artificial Intelligence in this topic.

We propose an approach for the design of multiagent simulations of human societies, in which the regulatory system is both the core and the main research topic of the simulation. The possibility to play with social norms in a multiagent system based on the simulation will help us understand how human societies processes react, answering questions such as:

1. What would be the most efficient set of norms to meet the target of demographic sustainability.
2. How far is the actual system from the optimal. And lastly to explain why can a normative system be so resilient despite the distance from the optimal.

2.1 Steps

2.1.1 Convening a common language

One of the most challenging previous tasks for both partners, the social scientist and the computer scientist, was the establishment of a common language. We stated that the standardisation of the natural language in different disciplines has developed gradually a bias in the meaning of those concepts used normally. Apart from this problem, there is another set of procedures, epistemological assumptions and conventions that must be understood by both parties. This is not a trivial effort. If you do not build a good starting the final result can be frustrating for both partners, as we have stated in other attempts of previous researches.

2.1.2 Sorting out the social theories to construct the basis of the model

The next step has been to find and select the definitions and social theories of those characters, processes and social relationships and behaviour that have had greater consensus. We need a number of starting points and anchors to launch our verification and also for calibrating the vectors, algorithms and formulations that we will use in the construction of the model and the programming of the agents characters and their behaviour. Although a priory it was thought that this would be a relatively easy step, this has not been the case. The fragmentation of the social sciences makes it difficult to look at such diverse sources to synthesize a series of axiomatic starting points completely consensual. Furthermore, the definition of what makes us basically humans also falls in the field of life sciences, which have failed to act either experimentally to verify their preconceptions. So the same parameters we use in the construction of the system have themselves become subject of experimentation.

2.1.3 Defining agents, variables and their life course

Using a set of parameters perfectly defined and identified in the literature, we defined our agents through a series of variables and states they go through along their life. These characters basically refer to those variables that are significant in making decisions and launch actions related to social reproduction. We set a general environment: a network composed of different levels, from the unit of close reproduction (a family) to the whole reproduction network (from an ethnical entity) and a general character of the society that will be distributed among the agents (i.e. rate of accidents, health, gender, mortality and fertility rates...).

2.1.4 Sorting the main social rules in actual hunter-gatherer societies and formalising them

To start our realistic simulations we have chosen four sub-actual hunter-gatherer societies, ethnographically well documented. These have been chosen taking into account the variability in the natural environment and the social complexity. We have compiled all the relevant ethnographic information and selected all the explicit rules of behaviour, social fables and morals that may affect social reproduction. Once sorted by their application to different segments of society, status, age and gender we have simplified them so they can be represented using a formal language.

2.1.5 The computational model

In contrast with more traditional archaeology simulators focused on resources and their management – our focus is on interactions among individuals and the regulation of those interactions through norms. In our approach the normative system establishes what an agent should and shouldn't do but, at the same time, an agent is free to follow or not the norms according to its personal goals.

The behaviour of an agent is determined by its current goals, its internal state, a set of social norms that regulates a context and its willingness to observe those norms. While the goals and the internal state are specific of each agent, the norms that regulate the behaviour are common and assumed to be known for all the agents in the simulation.

A norm in our simulator has a set of antecedents and a set of consequents. There can be two types of antecedents in a norm: facts about the internal state of the agent or the relationships of the agent with other members of the society (for example,“age$<$1”) and actions that have to be performed so the norm is activated (for example go-hunting). If all the antecedents are satisfied (in our example that the agents age is below 13 and the agent decides to go hunting), the consequents reflect: (i) how the internal state of the agent will change and (ii) if there are some actions that will be performed as a consequence. The norm can have also consequents that will become active if the norm is not observed. The actions in the consequents will induce new changes (on top of those associated directly to the norm) in the internal state of the agent once they are performed. An example of a norm is: If a man is married and his wife has a very low prestige level the man can divorce. In that case the woman will fall into disgrace. If the man does not divorce he will lose credit in front of the other members of the society

This norm can be formalised as:

```
Antecedents:
   - facts: man(X), woman(Y), married(X,Y),
            prestige(Y)<low
   - actions: divorce(X,Y))
```

```
If observed:
    delete(married(X,Y)), prestige(X)=, prestige(Y)-
If not observed:
    prestige(X)-
```

Where prestige(X)–, prestige(X)= means that the prestige of the individual will decrease or remain equal respectively.

That is, if there is a man and a woman, they are married (as reflected by the social network), the prestige of the woman is very low and the man decides to divorce, then the married relation is removed from the social network, the prestige of the man remains untouched and the prestige of the woman decreases even more. If the norm is not observed (the man decides not to divorce) then the prestige of the man decreases.

The norms are organised in normative levels. We distinguish three different normative levels:

Basic level. Here we find all the norms and rules dictated by the nature of the individual. Two types of norms/rules are found at this level: biological rules like, for example, A woman do not become fertile till she has the first menstruation and basic social norms, that although are not biological we assume are also part of the nature of the individual. The norms/rules at this level have only facts as antecedents and therefore the agent cannot influence on their activation. However, as we will see, the agent can decide to follow norms that belong to higher normative levels that can cancel the activation of certain norms/rules at this level.

Social level. The norms at this level are norms dictated by the society as a whole. There is no central authority or institution that imposes their observance but following or not one of these norms usually has implications in terms of how the individual will be considered among the other members of the society. The social position of an individual influences the kind, frequency and quality of interactions she can have.

Institutional level. Finally, at this level we find those norms dictated by central authorities and institutions. Apart from the social consequences in front of the rest of the society, not following one of these norms normally imply sanctions coming from the central authority. Although this level is not relevant for the current objective (we are dealing with societies without central authorities), it has been included for completeness and for future uses of the model.

Norms in the basic level define the default behaviour of the agent. The social and institutional levels modulate this default behaviour by reinforcing or restricting specific conducts. In our model an individual can decide to follow or not the norms in the social and institutional levels and by so doing, modify the default behaviour.

In addition to the three constructs just mentioned the state of the internal variables of an agent, its personal goals and the normative system there is a fourth element that determines the behaviour of an individual in our model: the social relationships. We assume that all the members of the society know about these social networks.

An agent in the simulator is defined by a set of internal variables that describe the state of the agent at each simulation step. Agents also have personal goals and satisfying those goals is their ‘raison d’être’. Each goal has an associated strength that represents the relevance of that goal for the agent.

The agents can perform actions, and these actions lead them to follow (or not) a norm by satisfying its antecedents. The set of possible actions is a closed set defined in each specific simulation scenario. We use the symbol ¬ to denote the opposite conduct associated to that action. For example we can have the action go-hunting and also the action ¬go-hunting. In the second case, the action the agent is taking is avoid to go hunting (whatever this means in that context).

Of course, the observance of norms has consequences for the agents. Every time the agent is in the dilemma of deciding if it is worth it or not to follow a norm, it analyses (by looking at the consequents of the norm) how the observance of that norm favors its personal goals. According to that, it takes the actions associated to follow or avoid the norm. Notice that if, for example, following the norm requires (as stated by the antecedents) go-hunting and the agent decides not to observe the norm, this implies that the agent will perform the action ¬go-hunting.

It can happen that following a norm favours the achievement of a specific goal but at the same time is in detriment of achieving another one. The (normalised) strength of each goal becomes the probability that the agent decides to follow the norm or not (and therefore favors some goals and disfavours others). The same principle is applied if there is more than one goal affected by the norm.

In each step of the simulation, the system evaluates for each agent what are the norms (in the three normative levels) that given the current internal state of the agent are candidates to be fired. For those candidate norms that have actions in their antecedents, the agent decides if it wants to perform the actions and, as a consequence follow the norm, or ignore those actions (so the norm is not observed). The result of the previous process is the set of norms that are candidate to be fired.

2.2 The YamanaSim

The YamanaSim system, depicted in figure 1, is composed of three major components: a Simulator Initializer, a Multiagent System (MAS) and a Rule engine.

The job of the Simulator Initialiser is to load a simulation specification file and setup the MAS and the rule engine accordingly with their initial values. The simulation specification file allows the user to define: the population of agents that will participate in the simulation, parameters to simulate the population dynamics, and the set of rules that will lead the agents actions. The population of agents can be defined in two ways: (i) by declaring all the agents inline where all the agents and its relationships (networks) are defined one by one in the configuration file or (ii) by using demographic population information. By using demographic population information the user can define large sets of agents population easily, although at the cost of losing some detail.

The MAS is in charge of the agent population, the social networks and the control of the simulation. The MAS component is built using Repast Simphony (http://repast.sourceforge.net/) a well known agent-based modelling toolkit. The agents in the multiagent system are instantiated following the directives of the Simulator Initialiser. An agent in the YamanaSim simulator has three major elements: a set of attributes, a set of goals to maximise or minimise, and a decision making module. The agents attributes, as we have seen before (see Figure 1) are used to store data like gender, age, health and prestige. The goals define the current objectives of the agent, and can change along time. Finally the decision-making module uses these goals to select the actions that will be performed by an agent.

Also part of the MAS component are the social networks. There can be multiple networks to define different relationships between agents e.g. family, kinship, dominance relationships and so on. These networks are also initialised by the Simulator Initializer and evolve along the simulation execution.

Finally, the Rule Engine is in charge of evaluating, every time step and, from an individual point of view, the set of rules that concern the agent. As we said, for each candidate rule, the rule engine asks

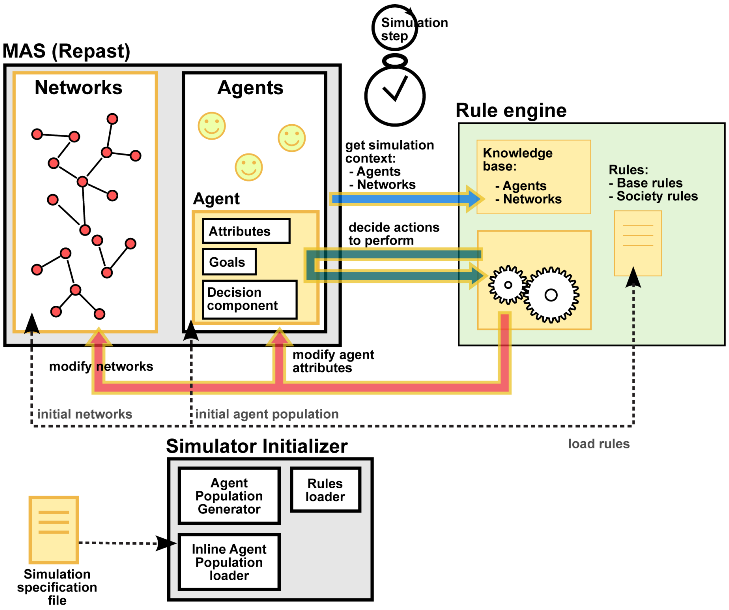


Figure 1. General schema of the architecture.

the agent about the actions to be performed (sending previously all the necessary information about the rule to the agent so the agent can reason about the rule and its consequences). This determines if the rule is finally fired or not.

3 Future work

We are currently running to test the functioning of a society (mostly biological) with a minimum of social norms and simplified agents. We will experience the development initially starting from a breeding pair, and then starting up from a band of 30 individuals. Later we will begin to add social norms and complexity to the agents and their behaviour, testing the effect of the introduction of such norms and variables to the demographic development of the artificial societies.

ACKNOWLEDGEMENTS

This work was supported by the CSIC [Proyecto Intramural de Frontera MacNorms, PIFCOO-08-00017]; the Generalitat de Catalunya [2005-SGR-00093] and the project “Analisis Etnoarqueologico de la reproduction en sociedades cazadoras-recoletoras” MINIECCO [HAR2011-24356].

REFERENCES

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

[2] David de la Cruz, Jordi Estevez, Pablo Noriega, Manuela Perez, Raquel Pique, Jordi Sabater, Assumpcio Vila, and Daniel Villatoro, ‘Nornas en sociedades cazadoras-pescadoras-recolectoras. argumentos para el uso de la simulacin social basada en agentes.’, *Cuadernos de prehistoria y arqueologa de la Universidad de Granada*, (20), 149–161, (2010).

[3] J. Estevez and et al., ‘Cazar o no cazar: es esta la cuestion?’, *Boletin de Antropologia Americana*, (33), 5–24, (1998).

[4] J. Estevez and A. Vila, *Twenty years of Ethnoarchaeological research in Tierra del Fuego: some thoughts for European Shell-Midden Archaeology*, 183–195, Oxbow Books, 2007.

[5] Assumpcio Vila, Jordi Estevez, Daniel Villatoro, and Jordi Sabater-Mir, ‘Archaeological materiality of social inequality among hunter-gatherer societies’, in *Archaeological Invisibility and Forgotten Knowledge Conference Proceedings (2007)*, volume BAR-S2183, pp. 202—210. Archaeopress, (2010).

Breaking immersion by creating social unbelievabilty

Henrik Warpefelt[1] and **Björn Strååt**[2]

Abstract. For the last 20 years, computer games and virtual worlds have made great advances when it comes to audiovisual fidelity. However, this alone is not sufficient to make the games seem believable – the game world must also seem to be alive. In order to accomplish this, the world must be populated by realistic characters who behave in a coherent and varied way. Many game developers seem to realize this, and the capacity of the artificial intelligence controlled non-player characters in the games are often large selling points. However, as pointed out by recent research these opponents do not always exhibit realistic, coherent and varied behaviour. We have examined this phenomenon by analysing a number of games where non-player characters are especially important for the players' enjoyment, and established six anti-heuristics that can be used to identify non-desirable behaviour in non-player characters.

1 Introduction

While many of today's games focus on the aural and visual experience of game play, some researchers have put forward that AI (Artificial Intelligence) may be bring games to a new level. Castronova [5] states that "*of all the technological frontiers in world-building, artificial intelligence (AI) holds the most promise of change*". A similar reasoning is presented by Bartle [1] who writes about the potential of AI-controlled non-player characters (NPCs) in games: "*from the point of view of world design, AI promises great things. If virtual worlds could be populated by intelligent NPCs, all manner of doors would open*". These two quotes both deal with the importance of NPCs in making the world feel alive.

The main factor in making the world feel alive is immersion - which is described by Bartle [1] as "*the sense that a player has of being* in *a virtual world*". If the player cannot immerse himself in the world and forget outside distractions the magic circle of the game, as described by Huizinga [6], collapses and the player's lusory experience is lessened. As such it is important that the NPCs act in such a way that they are seem to be creating a living world. This requires that the NPCs have varied and believable behaviours. If they do not, the player will soon begin to see patterns in how the NPCs act - as explained by Johansson & Verhagen [7]. If the player can see the proverbial clockwork ticking away in the background the player's immersion disappears and the magic circle is dispelled.

It should, however, be noted that some game developers have attempted to introduce more complex behaviours in their games. Famous examples of this include the game *Half-Life* [10] which at the time of launch received praise for the teamwork performed by the enemy NPCs present in the game. This trend was continued in *F.E.A.R* which utilized a goal-oriented architecture to further advance the teamwork capabilities of the NPCs, as explained by Orkin [9]. A more recent example is *Skyrim* [3], where the player's decisions have lasting effects on the world. These games all exhibit fairly complex social behaviours in the teamwork of the NPCs and their interaction with the players, but as Johansson & Verhagen [7] point out one can still see patterns in the behaviour of these NPCs.

However, in order to rectify any problems associated with repetitive behaviour in these NPCs we first need to describe these behaviours and make them explicit. This study aims to identify the types of NPC behaviour that has an adverse effect on the player's sense of immersion.

The next section (2) explains the pedigree of the method used in this article, as well as the preceding studies performed in this area. The changes made to the method in regard to preceding studies is explained in section 3. Section 4 explains how the games included in the study were selected. In section 5 the main work and data collection of this article are presented, followed by the resulting anti-heuristics in section 6 and the conclusions and future work in section 7.

2 Previous work

Johansson & Verhagen [7] used an adapted version of the *Carley & Newell Fractionation Matrix* (C&N matrix) presented by Carley & Newell in [4] to describe the attributes of their suggested architecture for more believable NPCs - the *Model Social Game Agent* (MSGA). Carley & Newell originally combined theories from sociology relating to human behaviour into a matrix to visualize what they call a Model Social Agent (MSA) – an agent with strong, human-like, social behaviour.

In the matrix (our adapted version can be seen in figure 1) the X-axis illustrates Knowledge; i.e. an agent's needed knowledge in relation to an increasingly advanced situation, going from non-social tasks (e.g. cutting wood) where the demand on knowledge is rather low and the agent can act on perfect information, to more refined cultural behaviour and perspective (e.g. upholding norms) where the agent acts on imperfect information. The Y-axis illustrates Processing, where the agent goes from being omniscient/omnipotent (OA, top left) with no need to gather information or reflect on its tasks, to an emotional cognitive agent (ECA, bottom row) who, in theory, lets "*emotions modify and limit the behaviour of the cognitive agent*" [4]. The further down and to the right one looks in the matrix, the more human-like the agent becomes. In total, the matrix contains 74 examples of social behaviour, such as "goal directed" and "crisis response".

A study similar to this one has been undertaken before by Lankoski & Björk [8]. Unfortunately, that study is based on a limited data set (one game) and used design patterns to examine one single NPC. The results of Lankoski & Björk's study are valuable,

[1] Department of Computer and Systems Sciences, Stockholm University, Sweden, email: hw@dsv.su.se

[2] Department of Computer and Systems Sciences, Stockholm University, Sweden, email: bjor-str@dsv.su.se

but may be hard to apply by the members of the game development industry who actually implement the NPC AI in games since they are of a rather abstract nature.

In a previous study, described in [11], we used a combination of Johansson & Verhagen's and Carley & Newell's matrices, resulting in a matrix with a total of 80 values, as seen in figure 1. This was a pilot study performed in order to examine the matrix's viability as an evaluation tool for NPC social capability in games.

3 Method

The methodology of this study is derived from our previous study [11] and as such uses our adapted version of the C&N matrix (once again found in figure 1). For the study described in this paper we clearly defined a sentence describing the meaning of each value in the matrix, and then used these definitions[3] when analysing the included games. Similarly to our approach in our previous study, we have taken a "black box" approach to the analysis of NPCs – we simply accepted the behaviour of the NPCs at face value rather than to try to understand what the actually programming was telling the NPC to do, much like a player without knowledge of game development would. The opposite of this would be to take a "glass box" approach and study the working innards of the NPCs. However, the "glass box" approach was discarded since it was considered unfeasible to persuade a large number of game developers to share their proprietary code with us.

The data for this study was collected by recording game play in 14 games (see table 1 below for a list of games) as video, continuing until no new behaviours were exhibited by the NPCs in the game. We observed a wide variety of situations, encompassing all kinds of social behaviour – ranging from street conversations to combat. These videos were then analysed in two stages, separated by 6 weeks and performed by a two researchers per study who strived for consensus. The approach of doing two separate studies, each using multiple researchers, was taken in order to ascertain the validity of the study by applying multiple layers of triangulation. During the analysis process, each scenario encountered in the videos was described in text and evaluated for possible immersion-breaking behaviour according to the values in the C&N matrix. Each value was considered separately for a given scenario.

The intermediate data created in the previous step was then used to determine the most significant values in the C&N matrix, more specifically the values that were violated in at least 5 games during either study. At this stage, a game's violation in regards to a value was only counted once. Hence, if a game violated a certain value seven times it was still only counted once for the purposes of selecting values. The significant values can be seen in table 2.

Lastly, the descriptions of the violations of the significant values were examined in order to find similarities between them. The similarities were then reformulated into a set of heuristics that can be used to evaluate how flawed an NPC is.

4 Included games

The games included in this study can be seen in table 1. These games were selected based on the following criteria:

- AAA-titles, i.e. big-budget studio titles
- The player takes the role of a single character at a time (but may have several helpers)
- Not older than 10 years

These critera were chosen in order to ensure that the games incorporate fairly recent technology, and had the necessary funds to actually put money into the development of the AI controlling NPCs within the game. We decided not to search among lower budget or independent titles, since they are less widespread, and thus, possibly, making any problems found less general. The list of games (see table 1) is a rough estimate on high end titles over the selected time frame, and may therefore not give an accurate estimate of the fidelity of lower end titles. However, the assumption is that problems found in high end titles will also be applicable to lower end titles, whereas the reverse may not be true.

Lastly we chose to limit ourselves to games where the player controls one character at a time, so that the representation of personal interaction would be easily recognizable and as such less prone to misunderstanding.

5 Applying the matrix

The end result of this research was a collection of heuristics, totaling a number of 6, based on an analysis of the 14 games (see table 1). These defined how NPCs should act if they intend to break the player's feeling of immersion, and are as such called "anti-heuristics".

The intended use for our anti-heuristics is reminiscent of the analysis of an NPC done by Lankoski & Björk [8] in Bethesda's role-playing game *Oblivion* [2], where the authors used different patterns to identify weaknesses in NPC behaviour. While Lankoski and Björk used their patterns to describe the *capabilities* of a given NPC, our anti-heuristics are intended to be used in identifying the *failings* of a given NPC. The reason for this reversed use is that it lets the analyst look for things the NPC *does* rather than things it *does not* do.

In doing the analysis of the games we found that certain values in the Carley & Newell fractionation matrix were more commonly occurring than others when the games broke the player's sense of immersion. These were *Adaption*, *Lack of Awareness*, *Models of Others* and *Models of Self*. The specific number of occurrences can be seen in table 2. These values display certain common traits between the situations we encountered in the games, and examples thereof are described in the sections below, along with the definitions we used for each value. While these values are not the only ones that were relevant to our study, they were by far the most commonly occurring.

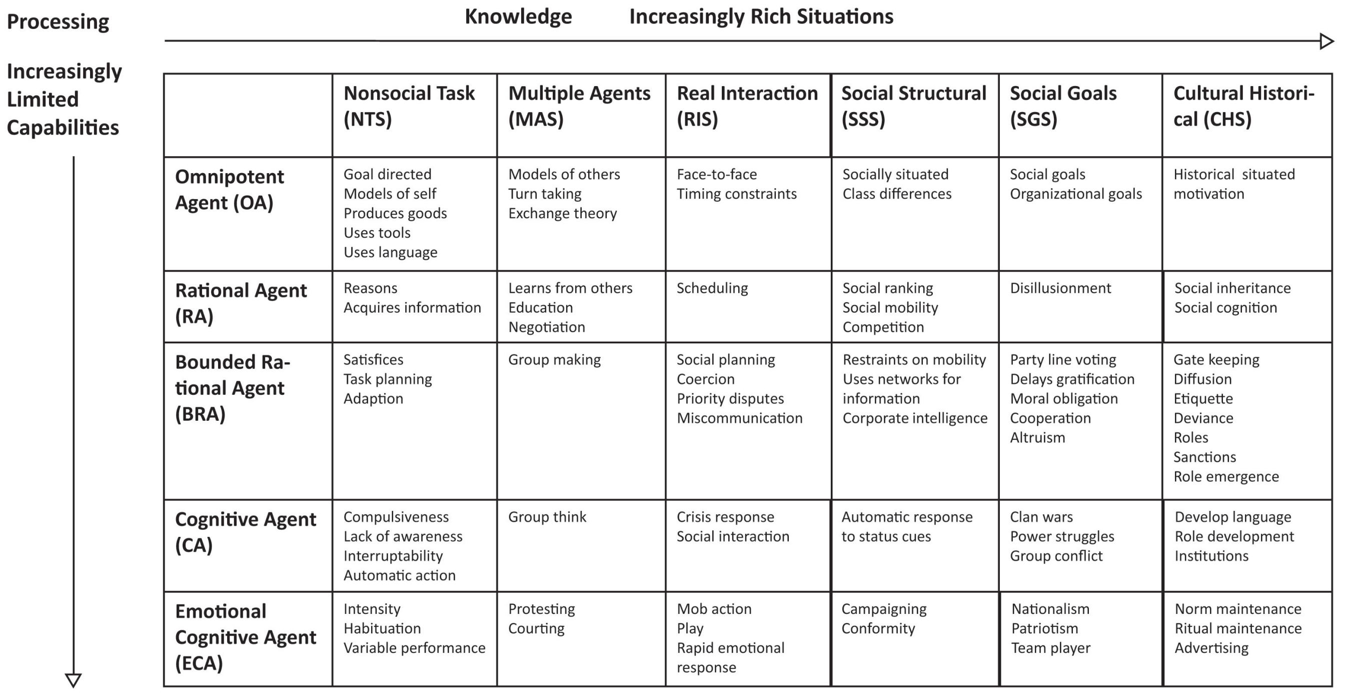
5.1 Adaption

For the purpose of this study we defined Adaption as "Characters adapt their behaviour to the present situation, including interrupting current tasks if the change in situation requires it". In the adapted C&N matrix, Adaption is located in the intersection of *Nonsocial Task* and *Omnipotent Agent*.

The analysis of the material in our data collection gives us two versions of Adaption failure; an entity either fails to adapt to improve its situation or adapts in such a way that its situation gets worse.

We obeserved an example of adapting to a worse situation in LA Noire. In this situation the player is inside a warehouse, engaged in a firefightand using a cupboard as cover. Hiding behind another cover a few meters in front of the player are two NPC gunmen, who take potshots at the player.

[3] These definitions have not been included in this paper since they would take up too much space, but the ones that are relevant to our results are presented as needed.



Processing
Knowledge
Increasingly Rich Situations
Increasingly Limited Capabilities

	Nonsocial Task (NTS)	Multiple Agents (MAS)	Real Interaction (RIS)	Social Structural (SSS)	Social Goals (SGS)	Cultural Histori-cal (CHS)
Omnipotent Agent (OA)	Goal directed Models of self Produces goods Uses tools Uses language	Models of others Turn taking Exchange theory	Face-to-face Timing constraints	Socially situated Class differences	Social goals Organizational goals	Historical situated motivation
Rational Agent (RA)	Reasons Acquires information	Learns from others Education Negotiation	Scheduling	Social ranking Social mobility Competition	Disillusionment	Social inheritance Social cognition
Bounded Ra-tional Agent (BRA)	Satisfices Task planning Adaption	Group making	Social planning Coercion Priority disputes Miscommunication	Restraints on mobility Uses networks for information Corporate intelligence	Party line voting Delays gratification Moral obligation Cooperation Altruism	Gate keeping Diffusion Etiquette Deviance Roles Sanctions Role emergence
Cognitive Agent (CA)	Compulsiveness Lack of awareness Interruptability Automatic action	Group think	Crisis response Social interaction	Automatic response to status cues	Clan wars Power struggles Group conflict	Develop language Role development Institutions
Emotional Cognitive Agent (ECA)	Intensity Habituation Variable performance	Protesting Courting	Mob action Play Rapid emotional response	Campaigning Conformity	Nationalism Patriotism Team player	Norm maintenance Ritual maintenance Advertising

Figure 1: The adapted Carley & Newell Fractionation matrix

Table 1: Games included in the study, sorted by title

Title	Developer	Year	Description
Assassin's Creed: Revelations	Ubisoft	2011	Historical fiction action role playing game
Dragon Age: Origins	Bioware	2009	Fantasy role playing game
Dragon Age 2	Bioware	2011	Fantasy role playing game
Fable 3	Lionhead Studios	2011	Fantasy role playing game
Fallout 3	Bethesda Softworks	2009	Postapocalyptic role playing game
Mass Effect	Bioware	2007	Science fiction action role playing game
Mass Effect 3	Bioware	2012	Science fiction action role playing game
L.A. Noire	Team Bondi/Rock Star Leeds	2011	Modern-day murder mystery game
RAGE	id Software	2011	Postapocalyptic first person shooter
The Elder Scrolls III: Morrowind	Bethesda Softworks	2002	Fantasy role playing game
The Elder Scrolls IV: Oblivion	Bethesda Softworks	2006	Fantasy role playing game
The Elder Scrolls V: Skyrim	Bethesda Softworks	2011	Fantasy role playing game
Vampire, the Masquerade: Bloodlines	Troika Games	2004	Fantasy role playing game
Warhammer 40,000: Space Marine	Relic Entertainment	2011	Science fiction third person shooter

Table 2: Significant values from both rating sessions

Value	Cell in C&N matrix	Study 1	Study 2
Adaption	NTS/BRA	4	9
Lack of Awareness	NTS/CA	6	7
Models of Others	MAS/OA	10	10
Model of Self	NTS/OA	2	8

When the camera is panned around, a third NPC gunman can be seen standing at a flank position, enabling him to shoot the player in the back. Instead of firing his weapon, the gunman runs away, into the player's field of fire and takes cover with his companions.

It can be posited that a person's rationality wavers when in a close quarter fire fight, but it seems very peculiar that a gunman should leave a good flanking position to put himself in a more risky position. The running gunman is a good example of Adaption failure.

5.2 Lack of Awareness

Lack of Awareness was defined as "Characters are unaware of events, not necessarily caused by other characters, happening in their immediate vicinity". In the adapted C&N matrix, Lack of Awareness is located in the intersection of *Nonsocial Task* and *Cognitive Agent*.

There are two sides to lack of awareness, over-awareness and obliviousness. The previous case, over-awareness, is illustrated by a scenario from *Vampire, the Masquerade: Bloodlines*. In this scenario the player walks through a house full of NPC thugs, who ignore him. He then walks out to the back of the house without anyone seeing him and shuts off the power. The thugs in the house are instantly aware of where the player is, that it was he who shut off the power and that they should attack him. Here the thugs show an extraordinary level of awareness - no one saw the player flip the power switch and yet they know that it was the player who shut off the electricity.

The latter case, obliviousness, is illustrated by a number of villagers in *Fable 3*. In this scenario the player is walking around in a village and shooting her pistol at the local NPC villagers, and since Fable 3 has a "safety mode" that can be turned on and off the player is currently unable to harm the villagers and the shots go slightly to the side. However, the villagers do not react in the slightest to the bullets flying around their heads but instead go about their daily business as if nothing had ever happened.

5.3 Models of Others

Models of Others is defined as "Characters are aware of what other entities are doing and where they are located". In the adapted C&N matrix, Adaption is located in the intersection of *Multiple Agents* and *Omnipotent Agent*.

We can exemplify Models of Others with a scenario from Skyrim: Here the player encounters a "fugitive" who hands over an item and tells the player to keep it safe, and that he will kill the player if she tells anyone. Any further interaction with the fugitive is fruitless; he simply repeats his former threats.

As the player follows the fugitive, they come upon a small pond in the forest, where a pair of hostile monster crabs reside. As they approach the pond, a hunter runs up to the player, and the fugitive cries for help and runs away to hide. The hunter engages the player in conversation, asking if she has seen a fugitive nearby, even though the fugitive just ran past the hunter, passing in plain sight no more than a few meters from him. The hunter even formulates his question as "Did you see anyone run past just now?". However, in approaching the player, the hunter is attacked by one of the monster crabs. The hunter questions the player about the fugitive while the crab happily gnaws away on the hunter. After the player finishes the conversation, the hunter dies from the crab attack.

The breaches of Models of Others here is that the hunter fails to observe the fugitive passing by and the monster crab attacking him.

5.4 Model of Self

Model of Self was defined as "Characters are aware that they are being affected by events happening around them". In the adapted C&N matrix, Adaption is located in the intersection of *Nonsocial Task* and *Omnipotent Agent*.

This is a scenario from Oblivion, where the player approaches a lizardman (an anthropomorphic lizard standing on two legs) standing close to the castle moat. By walking into the lizardman, the player is able to nudge him over the edge, into the moat. The lizardman falls in and starts treading water, with his head just over the surface, without complaining. The player jumps into the water and is merrily greeted by the lizardman. This lizardman seems unaware of what the player just did and is seemingly oblivious to his situation.

6 Anti-heuristics

After analysing a number of situations in the aforementioned games, we discovered a number of common weaknesses. These have been aggregated as a number of rules for how to *not* design an NPC. These are not solely based on the examples given above, but rather use a bigger data set. However, the examples above were partly chosen since they show these behaviours very well.

Our anti-heuristics are:

1. Ensure that the NPC always knows everything that is happening in the world. It should be omniscient!
2. Ensure that the NPC is seemingly unaware of things that it should feasibly be aware of.
3. Ensure that the NPC is seemingly unaware of what others are doing that could affect the NPCs, its friends or the environment.
4. Ensure that the NPC is seemingly unaware of actions performed that directly involve or affect it.
5. Ensure that the NPC always reacts in such a way that it makes its present situation worse.
6. Ensure that the NPC, through lack of reaction, never improves its situation.

These anti-heuristics may seem to cover overlapping areas, but that is wholly intentional. The rules are intended to be used together, and as such each individual rule does not contribute a lot of new knowledge. In this case, the total really is more than the sum of its parts.

The astute reader will also notice that they sometimes contradict each other, and this is also an intentional aspect of them. In the case of awareness, there seems to be a certain part of the awareness spectrum that makes a character seem believable.

7 Conclusion and final thoughts

The breaches of immersion that were found were frequent and often very obvious, and they occurred in a wide variety of games (representing roughly ten years of development). Given that some of our findings are seemingly obvious, and were found by simply playing the game, the developers cannot be unaware of these issues. While we understand that there are practical and financial limitations on how much effort can be put into creating believable NPC in games, but we have chosen not to include this consideration in our work. The goal was, as mentioned, to find NPC behaviours that negatively affect immersion, without taking any heed to the underlying system.

However, the issue of cost and commitment of the game developer is interesting. If we can communicate the results of this study to a game developer, and get a different perspective, we could certainly get material for several interesting reports.

Using this fairly straight forward and simple method we were able to isolate several immersion breaking errors, and while we presented a small set of anti-heuristics this list could easily be expanded using the same method. These anti-heuristics (ours and potential new ones) can then be of use to both researchers and game developers, since they allow for quick identification of potentially immersion-breaking situations.

Unfortunately the method used in this study is not without flaws. The C&N matrix is a rather unwieldy construct, and many of the values are less usable for games; examples of this would be values that imply an insight into the inner workings of the mind of an NPC, such as *Social Cognition*, which conflicts with our “black box” approach to the inner workings of the NPC. Other values that could have been of use were lacking in the matrix, such as the ability to evaluate if the way the NPCs navigate the world was believable. In order to remedy this issue we have begun work on a replacement model based on the C&N matrix, but better adapted to the context of computer games.

REFERENCES

[1] Richard Bartle, *Designing virtual worlds*, New Riders Publishing, 2004.

[2] Bethesda Softworks. The elder scrolls iv: Oblivion, 2007. Computer game.

[3] Bethesda Softworks. The elder scrolls v: Skyrim, 2011. Computer game.

[4] Kathleen Carly and Allen Newell, ‘The nature of the social agent’, *The Journal of Mathematical Sociology*, **19**(4), 221–262, (1994).

[5] Edward Castronova, *Synthetic Worlds: The Business and Culture of online games*, University of Chicago Press, 2005.

[6] Johan Huizinga, *Homo Ludens: a study of the play-element in culture*, Beacon Press, 1955.

[7] Magnus Johansson and Harko Verhagen, ‘”where is my mind”- the evolution of npcs in online worlds’, in *ICAART 2011 - Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*, eds., Joaquim Filipe and Ana L. N. Fred, volume 2, pp. 359–364. SciTePress, (2011).

[8] Petri Lankoski and Staffan Björk, ‘Gameplay design patterns for believable non-player characters’, in *Situated Play: Proceedings of the 2007 Digital Games Research Association Conference*, ed., Baba Akira, pp. 416–423, Tokyo, (2007). The University of Tokyo.

[9] Jeff Orkin, ‘Three states and a plan: The a.i. of f.e.a.r.’, in *Game Developers Conference*, (2006).

[10] Valve Corporation. Half-life, 1998. Computer game.

[11] Henrik Warpefelt and Björn Strååt, ‘A method for comparing npc social ability’, in *Proceedings of the 5th Annual International Conference on Computer Games and Allied Technology (CGAT 2012)*, ed., Edmond Prakash, pp. 58–63. Global Science & Technology Forum, (2012).