AISB/IACAP World Congress 2012

Birmingham, UK, 2-6 July 2012

Social Computing, Social Cognition, Social Networks and Multiagent Systems
Social Turn - SNAMAS 2012

Gordana Dodig-Crnkovic, Antonino Rotolo, Giovanni Sartor, Judith Simon, and Clara Smith (Editors)

Part of

Published by
The Society for the Study of
Artificial Intelligence and
Simulation of Behaviour

http://www.aisb.org.uk

ISBN 978-1-908187-18-5

Foreword from the Congress Chairs

For the Turing year 2012, AISB (The Society for the Study of Artificial Intelligence and Simulation of Behaviour) and IACAP (The International Association for Computing and Philosophy) merged their annual symposia/conferences to form the AISB/IACAP World Congress. The congress took place 2–6 July 2012 at the University of Birmingham, UK.

The Congress was inspired by a desire to honour Alan Turing, and by the broad and deep significance of Turing's work to AI, the philosophical ramifications of computing, and philosophy and computing more generally. The Congress was one of the events forming the Alan Turing Year.

The Congress consisted mainly of a number of collocated Symposia on specific research areas, together with six invited Plenary Talks. All papers other than the Plenaries were given within Symposia. This format is perfect for encouraging new dialogue and collaboration both within and between research areas.

This volume forms the proceedings of one of the component symposia. We are most grateful to the organizers of the Symposium for their hard work in creating it, attracting papers, doing the necessary reviewing, defining an exciting programme for the symposium, and compiling this volume. We also thank them for their flexibility and patience concerning the complex matter of fitting all the symposia and other events into the Congress week.

John Barnden (Computer Science, University of Birmingham)
Programme Co-Chair and AISB Vice-Chair
Anthony Beavers (University of Evansville, Indiana, USA)
Programme Co-Chair and IACAP President
Manfred Kerber (Computer Science, University of Birmingham)
Local Arrangements Chair

Preface from the Symposium Chairs

Social Computing, Social Cognition, Social Networks, and Multiagent Systems (Social Turn - SNAMAS 2012), co-located in Birmingham (UK) with the AISB/IACAP World Congress 2012 - Alan Turing 2012, was organized to meet scholars working on social computing, i.e., the cross-fertilization between social science, philosophy, and computer science. This 2012 symposium merges the symposium *Social Turn: Social Computing - Social Cognition - Social Intelligence* and the SNAMAS symposium, focused on Social Networks and Multi-Agent Systems, which have earlier symposia in Social Computing at IACAP and the SNAMAS in AISB conferences.

The field of social computing has two aspects: the social and computational ones. There is the focus on socialness of social software or social web applications. Widespread examples of social software are blogs, wikis, social bookmarking services, instant messaging services, and social networking sites. Social computing often uses various types of crowdsourcing techniques for aggregation of input from numerous users (public at large). Tools such as prediction markets, social tagging, reputation and trust systems as well as recommender systems are based on collaborative filtering and thus a result of crowdsourcing. Another focus of social computing is on computational modeling of social behavior, among others through Multi-agent systems (MAS) and Social Networks (SN). MAS have an anchoring going beyond social sciences even when a sociological terminology is often used. There are several usages of MAS: to design distributed and/or hybrid systems; to develop philosophical theory; to understand concrete social facts, or to answer concrete social issues via modelling and simulation. MAS aim at modelling, among other things, cognitive or reactive agents who interact in dynamic environments where they possibly depend on each other to achieve their goals. The emphasis is nowadays on constructing complex computational systems composed by agents which are regulated by various types of norms, and behave like human social systems. Finally, Social networks (SN) are social structures made of nodes (which are, generally, individuals or organizations) that are tied by one or more specific types of interdependency, such as values, visions, idea, financial exchange, friends, kinship, dislike, conflict, trade, web links, disease transmission, among many others. Social networks analysis plays a critical role in determining the way specific problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals. Social networks analysis

has addressed also the dynamics issue, called dynamic networks analysis. This is an emergent research field that brings together traditional social network analysis, link analysis and multi-agent systems.

The contributions in these proceedings include two abstracts for the two invited keynote presentations and a selection of 20 papers addressing a wide range of topics, such as: Logical, Computational and Theoretical Models for MAS; Social Simulation: Theory and Practice; Trust & Responsibility; Agency & Sociality; Legal, Ethical and Philosophical Aspects of MAS; Networks and MAS: Experimental Results. The accepted papers were carefully selected after a rigorous peer-review process. We thank the reviewers for their effort and very valuable contribution; without them it would not be possible to maintain and improve the high scientific standard the symposium has now achieved. We thank the authors for submitting good papers, responding to the reviewers' comments, and abiding by our schedule. We thank the keynote speakers, Marek Sergot and Bernhard Rieder, for their interesting contributions and presentations. And we thank the AISB/IACAP World Congress 2012 organizers for enabling this fruitful collocation of our symposium.

Finally, we would like to express our gratitude to our sponsor, the European Network for Social Intelligence, whose financial support helped us to organize this event.

Gordana Dodig-Crnkovic (Mälardalen University, Sweden)
Antonino Rotolo (CIRSFID, University of Bologna, Italy)
Giovanni Sartor (EUI and CIRSFID, University of Bologna, Italy)
Judith Simon (University of Vienna, Austria and
Karlsruhe Institute of Technology, Germany)
Clara Smith (UNLP and UCALP, Argentina)

Symposium Organization

Chairs

Gordana Dodig-Crnkovic	(Mälardalen University, Sweden)
Antonino Rotolo	(CIRSFID, University of Bologna, Italy)
Giovanni Sartor	(EUI and CIRSFID, University of Bologna, Italy)
Judith Simon	(University of Vienna, Austria and Karlsruhe Institute of Technology, Germany)
Clara Smith	(UNLP and UCALP, Argentina)

Program Committee

Doris Allhutter
Frederic Amblard
Giulia Andrighetto
Carlos Areces
Guido Boella
Pompeu Casanovas
Cristiano Castelfranchi
Mark Coeckelbergh
Diego Compagna
Rosaria Conte
Hamid Ekbia
Charles Ess
Christian Fuchs
Ricardo Guibourg
Lars-Erik Janlert
Matthias Mailliard
Antonio A. Martino
Jeremy Pitt
Melina Porto
Leon Van der Torre
Serena Villata
Jutta Weber

External Reviewers

María Grazia Mainero
Migle Laukyte
Leandro Mendoza
Agustin Ambrossio

Contents

Action, Agency and Causation

Marek Sergot[1]

The following is an old puzzle concerning the notion of 'proximate cause' discussed in legal theory. It has several versions. Here is one. The specific details are not important.

A certain traveller must cross the desert. It is well known that an adult needs two goat-skins of water to survive the journey. In readiness for an early start the traveller packs his camels before going to sleep. During the night an enemy comes and replaces the water in the goat-skins with poison. Later a second enemy comes, and not knowing what the first has done, makes small pinholes in the goat-skins so that the contents leak out. In the morning the traveller sets off, but finding his goat-skins empty, he dies in the desert. Which of the two enemies, if either, killed him?

There are two cyclists speeding towards each other on a cycle path. If both swerve to the left they will avoid a collision. If both swerve to the right they will avoid a collision. Otherwise they will collide. Suppose one swerves to the left and the other swerves to the right. Which, if either, caused the collision? It seems quite wrong to pick on one or the other: they both, collectively, were responsible. Suppose that on another occasion (it is a dangerous path) a pedestrian steps out and forces one of them to swerve left just as the other chooses to swerve right. Who then was responsible for the crash?

It is forbidden for a man and a woman to be alone in a room. There are two men and one woman in a room. One of the men gets up and leaves the room leaving the other man and the woman alone. Which of them is at fault? The man who left? The man who stayed? The woman? All of them, collectively?

The logic of agency is concerned with expressions of the form 'agent x brings it about that A', or 'agent x is responsible for its being the case that A', or 'the actions of agent x are the cause of its being the case that A', or more generally, 'the actions of the set of agents G, collectively, are responsible for, are the cause of, its being the case that A'.

The study of such logics has a very long tradition. The best known examples are perhaps the 'stit' ('seeing to it that') family (see, e.g., [Belnap and Perloff 1988, Horty and Belnap 1995, Horty 2001, Xu 1998, Belnap et al. 2001, Balbiani et al.2008]). [Segerberg 1992] provides a summary of early work in this area, and [Hilpinen 1997] an overview of the main semantical devices that have been used, in 'stit' and other approaches. With some exceptions (notably [Pörn 1977]) the semantics is based on a branching-time structure of some kind.

I have been working on a formal language that combines a logic of agency with a transition-based account of action: the semantical framework is a form of labelled transition system extended with an extra component that picks out the actions of a particular agent in any given transition. There is a two-sorted modal language for talking about properties of states and about the actions of individual agents or groups of agents in transitions, including two defined modalities of the 'brings it about' kind. The account can be generalised to produce some characterisations of collective agency, that is, of expressions of the form 'the set G of agents, collectively though perhaps unwittingly, brings it about that A'. The formal framework has been implemented in the form of a model checker that can evaluate formulas expressing properties of interest on (a symbolic representation of) an agent-stranded transition system.

One important distinguishing feature is that the framework seeks to deal with unintentional, perhaps accidental or unwitting, action as well as deliberative, purposeful or intentional action. As [Hilpinen 1997] observes: "The expression 'seeing to it that A' usually characterises deliberate, intentional action. 'Bringing it about that A' does not have such a connotation, and can be applied equally well to the unintentional as well as intentional (intended) consequences of one's actions, including highly improbable and accidental consequences." The agency modalities are of this latter 'brings it about' kind.

This is for both practical and methodological reasons. From the practical point of view, there is a wide class of applications for systems composed of agents, human or artificial, with reasoning and deliberative capabilities. There is an even wider class of applications if we consider also simple 'lightweight' agents with no reasoning capabilities, or systems composed of simple computational units in interaction. I want to be able to consider this wider class of applications too. From the methodological point of view, it is clear that genuine collective or joint action involves a very wide range of issues, including joint intention, communication between agents, awareness of other agents' capabilities and intentions, and many others. I want to factor out all such considerations, and investigate what can be said about individual or collective agency when all such considerations are ignored. The logic of unwitting collective agency might be extended and strengthened in due course by bringing in other factors such as (joint) intention one by one; we do not discuss any such possibilities here.

The talk will sketch the main components of this formal framework, but it will concentrate on examples rather than technical details. I will show how it deals with examples such as those above, and others. I will also identify some inadequacies and directions for further work: I will try to identify, for instance, why the current version cannot deal with the traveller example (except in a rather surprising and unsatisfactory way).

[1] Department of Computing, Imperial College London, SW7 2AZ, UK. E-mail: m.sergot@imperial.ac.uk

REFERENCES

[Balbiani et al.2008] Philippe Balbiani, Andreas Herzig, and Nicolas Troquard. Alternative axiomatics and complexity of deliberative stit theo-

ries. *Journal of Philosophical Logic*, 37(4):387–406, 2008.

[Belnap and Perloff 1988] N. Belnap and M. Perloff. Seeing to it that: a canonical form for agentives. *Theoria*, 54:175–199, 1988. Corrected version in [Belnap and Perloff 1990].

[Belnap and Perloff 1990] N. Belnap and M. Perloff. Seeing to it that: a canonical form for agentives. In H. E. Kyburg, Jr., R. P. Loui, and G. N. Carlson, editors, *Knowledge Representation and Defeasible Reasoning*, volume 5 of *Studies in Cognitive Systems*, pages 167–190. Kluwer, Dordrecht, Boston, London, 1990.

[Belnap et al. 2001] Nuel Belnap, Michael Perloff, and Ming Xu. *Facing the future: Agents and choices in our indeterminist world*. Oxford University Press, 2001.

[Hilpinen 1997] R. Hilpinen. On action and agency. In E. Ejerhed and S. Lindström, editors, *Logic, Action and Cognition—Essays in Philosophical Logic*, volume 2 of *Trends in Logic, Studia Logica Library*, pages 3–27. Kluwer Academic Publishers, Dordrecht, 1997.

[Horty 2001] J. F. Horty. *Agency and Deontic Logic*. Oxford University Press, 2001.

[Horty and Belnap 1995] J. F. Horty and N. Belnap. The deliberative stit: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24(6):583–644, 1995.

[Pörn 1977] Ingmar Pörn. *Action Theory and Social Science: Some Formal Models*. Number 120 in Synthese Library. D. Reidel, Dordrecht, 1977.

[Segerberg 1992] K. Segerberg. Getting started: Beginnings in the logic of action. *Studia Logica*, 51(3–4):347–378, 1992.

[Xu 1998] Ming Xu. Axioms for deliberative stit. *Journal of Philosophical Logic*, 27:505–552, 1998.

The Politics of Formalization: What social Computing Can Learn from the Prehistory of PageRank

Bernhard Rieder[1]

Social networking sites, but also various other online services, such as search engines, rely on computational techniques, developed from the 1950s onward, to filter, rank, suggest or modulate visibility and navigational distance. Authority, reputation, and relevance are attributed by means of "mechanical reasoning", which often produces significant real-world consequences. Techniques based on social network analysis making use of graph theoretical methods are among the most common tools to establish such distinctions based on formal criteria. PageRank, a method for scoring documents in a hypertext database, has achieved particular prominence due to its use in the worlds most successful search engine.

This presentation will focus on this particular technique as a *pars pro toto* in order to examine the different levels of theoretical and exceedingly political *commitments* built into the algorithm. Instead of merely treating it as an atemporal procedure, I will summarily reconstruct its lineage and prehistory to show in which way particular representations of "the social", set in specific currents of sociological thinking, are formalized and made operational by the PageRank method. If we consider the systems making use of methods like this one to be, at the same time, *descriptive* and *prescriptive* devices that represent and intervene in processes of knowledge production and social interaction, the theoretical assumptions underlying efforts in formalization and modeling merit particular attention and critical scrutiny. In the case of PageRank, sociometry and social exchange theory provide the "epistemological support" for the formal model and a reconstruction of this particular historical and intellectual context provides not only a better understanding of what the algorithm actually *does*, but also of the inevitable political dimension attached to procedures that constantly *arbitrate* between actors and their accounts of reality by conferring visibility or "centrality" to some and not to others.

The goal of this exercise is to outline a mode of analysis of formal methods used in social computing that relies on a historical and conceptual approach to provide a set of *resources* for both the interpretation and assessment of these methods. This implies different levels of analysis. On a more general level, one can start from the observation that both sociometry and social exchange theory have been (sometimes strongly) criticized for the specific assumptions and choices they make, and this presentation will try to show how this critique can be made useful in the analysis of the PageRank model. On a more specific level however, one can examine specific elements of the model, in particular the "dampening factor" used to reduce the propagation of "status" in the hypertext network, from the perspective of the sociological theories in question and ask which kind of theoretical commitment they encode, in a very practical sense.

The increasing application of algorithmic techniques to the structuring of social relationships intensifies and highlights the political dimension of these techniques. This presentation aims at sketching one possible way to approach this problem by treating PageRank as social theory expressed in algorithmic form.

[1] University of Amsterdam.

Decidability via Filtration of Neighbourhood Models for Multi-Agent Systems

Clara Smith[1] and **Leandro Mendoza**[2] and **Agustín Ambrossio**[3]

Abstract. Lately, many multi-agent systems (MAS) are designed as multi-modal systems [9, 15, 23, 22, 26, 28, 18]. Moreover, there are different techniques for combining logics, such as products, fibring, fusion, and modalisation, among others [1, 14, 16]. In this paper we focus on the combination of special-purpose logics for building "on demand" MAS. From these engineering point of view, among the most used normal logics for modeling agents' cognitive states are logics for beliefs, goals, and intentions, while, perhaps, the most well-known non-normal logics for MAS is the logic of agency (and, possibly, ability). We explore combinations of these normal and non-normal logics. This lead us to handle Scott-Montague structures, (neighbourhood models, in particular) which can be seen as a generalization of Kripke structures [20].

Interested in the decidability of such structures, which is a guarantee of correct systems and their eventual implementations, we give a new presentation for existing theorems that generalize the well-known results regarding decidability through the finite model property via filtrations for Kripke structures. We understand that the presentation we give, based on neighbourhood models, better fits the most accepted and extended logic notation actually used within the MAS community.

1 Motivation and Aims

In [32] Smith and Rotolo adopted [13]s cognitive model of individual trust in terms of necessary mental ingredients which settle under what circumstances an agent x trusts another agent y with regard to an action or state-of-affairs, i.e. under which beliefs and goals an agent delegates a task to another agent. Using this characterization of individual trust, the authors provided a logical reconstruction of different types of collective trust, which for example emerge in groups with multi-lateral agreement, or which are the glue for grounding *in solidum* obligations raising from a "common front" of agents (where each member of the front can behave, in principle, as creditor or debtor of the whole). These collective cognitive states were characterized in [32] within a multi-modal logic based on [9]s axiomatisation for collective beliefs and intentions combined with a non-normal modal logic for the operator Does for agency.

In a subsequent work, the multi-relational model in [32] was reorganized as a fibring, a particular combination of logics which amounts to place one special-purpose normal logics on top of another [31]. In this case, the normal logic was put on top of the non-normal one. For doing this, authors first obtained two restrictions of the original logics. By exploiting results in regard to some techniques for combining logics, it was proved that [32]s system is complete and decidable. Hence, the sketch for an appropriate model checker is there outlined.

One motivation regarding a further combination of those special purpose logics for MAS is the aim to have an expressive enough system for modelling interactions between a behavioural dimension and a cognitive dimension of agents, and testing satisfiability of the corresponding formulas. For example, for modelling expressions such as $\mathrm{Does}_i\,(\mathrm{Bel}_j\,\mathscr{A})$ which can be seen as a form of persuasion or influence: agent i makes agent j have $\mathscr{A}$ as belief. This formula cannot be written in the fibred language in [31] neither in the original language in [32] because such languages have a restriction over the form of the *wffs*: no modal operator can appear in the scope of a Does. In [31], authors outlined a combination of the normal and the non-normal counterparts of the base logics. That combination lead to an ontology of pairs of situations allowing a structural basis for more expressiveness of the system. That combination is the result of (again) splitting of the original structure, which is a multi-relational frame of the form [32, 17]:

$$\mathfrak{F} = \langle A, W, \{B_i\}_{i\in A}, \{G_i\}_{i\in A}, \{I_i\}_{i\in A}, \{D_i\}_{i\in A}\rangle$$

where: A is a set of agents, W is a set of posible worlds, and $\{B_i\}, \{G_i\}, \{I_i\}, \{D_i\}$ are the accessibility relations for beliefs, goals, intentions, and agency respectively. The underlying set of worlds of the combination is an ontology of pairs of worlds (w_N, w_D). There are two structures where to respectively test the validity of the normal modalities and the non-normal modalities. The former is a Kripke model; the latter a neighbourhood model. The definition of a formula being satisfied in the combined model at a state (w_N, w_D) amounts to a scan through the combined structure, done according to which operator is being tested. Normal operators move along the first component w_N, and non-normal operators move along the second component of the current world w_D.

Regarding the application to agents, it is also common that the cognitive modalities are extended with temporal logics. For example, Schild [29] provides a mapping from Rao and Georgeff's BDI logic [27] to μ-calculus [24]. The model of Rao and Georgeff is based on a combination of the branching time logic CTL^* [8] and modal operators for beliefs, desires, and intentions. Schild collapses the (original) two dimensions of time and modalities onto a one dimensional structure. J. Broersen [5] presents an epistemic logic that incorporates interactions between time and action, and between knowledge and action.

Correspondingly, H. Wansing in [2] points out that (i) agents act in time, (ii) obligations change over time as a result of our actions and the actions of others, and (iii) obligations may depend on the

[1] FACET, UCALP, Argentina and Facultad de Informática, UNLP, Argentina
[2] Facultad de Informática, UNLP, Argentina and CONICET
[3] FACET, UCALP, Argentina

future course of events. In ([2], Section 10.3) he adopts a semantics reflecting the non-determinism of agency: models are based on trees of moments of time branching to the future. Agentive sentences are history dependent, formulas are not evaluated at points in time but rather at pairs (*moment*, *history*), where *history* is a linearly ordered set of moments.

Cohen and Levesque [7, 21] embed, using function mappings, a modal logic of beliefs and goals with a temporal logic with non-deterministic and parallel features.

In this paper we define a combination of logics for MAS as a special case of neighbourhood structures. Previously, we give a new presentation of decidability results which apply to a particular kind of models: neighbourhood models. In the literature, the analysis of transfer of logical properties from special purpose logics to combined ones is usually based on properties of normal logics. It is claimed that the proof strategies in the demonstration of transference of properties of normal logics could in principle be applied to non-normal modal logics [12]. In a mono-modal logic with a *box* modality, normality implies that the following formulas are valid: $\Box(p \rightarrow q) \rightarrow (\Box p \rightarrow \Box q)$ and $\Box(p \wedge q) \leftrightarrow (\Box p \wedge \Box q)$, as well as the admission of the rule from $\vdash \mathcal{A}$ infer $\vdash \Box\mathcal{A}$ [3, 12]. None of this is assumed to hold for a non-normal logics. We indeed use a non normal modal logic for agency, as developed by Elgesem [11, 17]; and aim to put it to work with normal logics for, e.g, beliefs and goals. The logic of agency extends classical propositional logic with the unary symbol Does satisfying the following axioms: $\neg(\mathrm{Does}\,\top)$, $(\mathrm{Does}\,\mathscr{A}) \wedge (\mathrm{Does}\,\mathscr{B}) \Rightarrow \mathrm{Does}(\mathscr{A} \wedge \mathscr{B})$ and $\mathrm{Does}\,\mathscr{A} \Rightarrow \mathscr{A}$ together with the rule of Modus Ponens and the rule saying that from $\mathscr{A} \Leftrightarrow \mathscr{B}$ you can conclude $\mathrm{Does}\,\mathscr{A} \Leftrightarrow \mathrm{Does}\,\mathscr{B}$. The intended reading of $\mathrm{Does}\,\mathscr{A}$ is that 'the agent brings it about that $\mathscr{A}$'. (See Section 2.1 in [11].) A detailed philosophical justification for this logic is given in [11] and neighborhood and selection function semantics are discussed in [11, 17].

One advantage regarding the choice of a logic of agency such as Does relies on the issue of action negation. For Does, and for other related logics of action such as the one in [5], action negation is well-understood: given that the logic for Does is Boolean, it is easy to determine what $\neg\mathrm{Does}\,\mathscr{A}$ means. This allows providing accurate definitions for concepts such as e.g. "refrain", especially useful in normative MAS: I have the opportunity and ability to do something, but I do not perform it as I have the intention not to. Up to now, although addressed, there are no outstanding nor homogeneous solutions for the issue on action negation in other relevant logics for MAS such as dynamic logics (see e.g. [4, 5, 25]).

We organize the work as follows. In Section 2 we directly adapt for neighbourhood models the strategy in [3] regarding the *finite model property (FMP) via filtration*. This includes: (i) establishing conditions for finding a filtration of a neighbourhood model, (ii) the demonstration of a filtration theorem for the neighbourhood case, (iii) guaranteeing the existence of a filtration, and (iv) the proof of the *FMP Theorem* for a mono-modal neighbourhood model. In Section 3 we show how the results in Section 2 can be applied for proving decidability of a neighbourhood model with more than one modality. We also devise examples for a uni-agent mono-modal non-normal system, a uni-agent multi-modal system and a multi-modal multi-agent system. In Section 4 we concentrate on a combined *MAS*, with an underlying neighbourhood structure. Conclusions end the paper.

2 Decidability for the neighbourhood case through the extension of the FMP strategy for the Kripke case.

We mentioned that normal logics can be seen as a platform for the study of transference of decidability results for non-normal logics and combination of logics. We rely on well-studied results and existing techniques for Kripke structures, which are usual support of normal logics, to provide a new presentation of existing decidability results for a more general class of structures supporting non-normal logics.

We start from the definitions given by P. Blackburn *et. al.* [3]. In [3](Defs. 2.36, 2.38 and 2.40), the construction of a finite model for a Kripke structure is supported in: (i) the definition of a filtration, (ii) the Filtration Theorem, (iii) the existence of a filtration for a model and a subformula closed set of formulas, and (iv) the Finite Model Property Theorem via Filtrations.

B. Chellas, in its turn, defined filtrations for minimal models in [6] (Section 7.5). Minimal models are a generalization of Kripke ones. A minimal model is a structure $\langle W, N, P\rangle$ in which W is a set of possible worlds and P gives a truth value to each atomic sentence at each world. N, is a function that associates with each world a collection of sets of worlds. The notation used throughout is one based on truth sets ($\|\mathscr{A}\|$ is the set of points in a model where the *wff* $\mathscr{A}$ is true). Truth sets are a basic ingredient of selection function semantics.

In what follows we give a definition of filtration for Scott-Montague models using a neighbourhood approach and notation. Neighbourhood semantics is the most important (as far as we consider) generalization of Kripke style (relational) semantics. The set of possible worlds is replaced by a Boolean algebra, then the concept of validity is generalized to the set of true formulas in an arbitrary subset of the Boolean algebra, but (generally for every *quasi-classical* logics) the subset must be a filter. This 'neighbourhood approach' focuses on worlds, which directly leads us to the underlying net of situations that ultimately support the system: relative to a world w we are able to test whether agents believe in something or carry out an action. The neighbourhood semantics better adapts to the specification of most prevailing modal multi-agent systems, which lately tend to adopt the Kripke semantics with a notation given as in [3]. This because, probably, that notation is more intuitive for dealing with situations and agents acting and thinking according to situations, rather than considering formulas as 'first class' objects. This is crucial in current practical approaches to agents; in a world an agent realises its posibilities of succesful agency of $\mathscr{A}$, its beliefs, it goals, all relative to the actual world w, In this perspective, situations are a sort of "environmental support" for agent's *internal configuration* and *visible actions*. Worlds are, therefore, in a MAS context, predominantly, abstract descriptions of external circumstances of an agent's community that allow or disallow actions, activate or nullify goals.

That is why we prefer to work with neighbourhood models as models for MAS, keeping in mind that, while it is possible to devise selection function models for MAS, this is not nowadays usual practice. Also, as it is well-known, the difference between selection function semantics and neighbourhood semantics is merely at the intuitive level (their semantics are equivalent, and both known as ScottMontague semantics [17]).

P. Schotch has already addressed the issue of paradigmatic notation and dominating semantics for modalities. In his work [30] he points out that the necessity truth condition together with Kripkean structures twistedly "represent" the model-theoretic view of the area, given that -among other reasons- many "nice" logics can be devised

with those tools. Moreover, due to this trend, he notes that previous complex and important logics (due to Lewis, or to the "Pennsylvania School") have become obsolete or curiosities just because their semantics is less elegant.

We adopt an eclectic position in this paper: we choose a structure that allows non-normal semantics and we go through it with the notation as given in [3], which is currently well-accepted and well-understood for modal MAS.

Next we outline some tools for finding a filtration of a neighbourhood model. We generalize the theorems for Kripke structures given in [3].

Definition 1 (Neighbourhood Frame)**.** A neighbourhood frame [20, 6] is a tuple $\langle W, \{N_w\}_{w\in W}\rangle$ where:

1. W is a set of worlds, and
2. $\{N_w\}_{w\in W}$ is a function assigning to each element w in W a class of subsets of W, the neighbourhoods of w.

We will be working with a basic modal language with a single unary modality, let us say '#'. We asume that this modality has a neighbourhood semantics. For example, '#' may be read as the Does operator, or an ability operator, as proposed by Elgesem [11]; or represent a "refrain" operator based on Does and other modalities such as ability, opportunity and intentions.

Definition 2 ((Recall Def. 2.35 in [3]) Closure)**.** A set of formulas Σ is closed under subformulas if for all formulas φ, if $\varphi \vee \varphi' \in \Sigma$ then so are φ and φ'; if $\neg\varphi \in \Sigma$ then so is φ; and if $\#\varphi \in \Sigma$ then so is φ. (For the Does modality, for example, if $\mathrm{Does}\,\varphi \in \Sigma$ so is φ).

Definition 3 (Neighbourhood Model)**.** We define $\mathfrak{M} = \langle W, \{N_w\}, V\rangle$ to be a model, where $\langle W, \{N_w\}\rangle$ is a neighbourhood frame, and V is a valuation function assigning to each proposition letter p in Σ a subset $V(p)$ of W (i.e. for every propositional letter we know in which worlds it is true).

Given Σ a subformula closed set of formulas and given a neighbourhood model $\mathfrak{M}$, let $\equiv_\Sigma$ be a relation on the states of $\mathfrak{M}$ defined by $w \equiv_\Sigma v$ iff $\forall\,\varphi \in \Sigma$ ($\mathfrak{M}, w \models \varphi$ iff $\mathfrak{M}, v \models \varphi$). That is, for all *wff* φ, φ is true in w iff is also true in v. Clearly $\equiv_\Sigma$ is an equivalence relation. We denote the equivalence class of a state w of $\mathfrak{M}$ with respect to $\equiv_\Sigma$ by $[w]_\Sigma$ (or simply $[w]$ when no confusion arises).

Let $W_\Sigma = \{[w]_\Sigma \,/\, w \in W\}$.

Next we generalize for neighbourhood models the concept of filtration given in [3].

Definition 4 (Filtrations for the neighbourhood case)**.** Suppose $\mathfrak{M}^f$ is any model $\langle W^f, \{N_w\}^f, V^f\rangle$ such that $W^f = W_\Sigma$ and:

1. If $U \in N_w$ then $\{[u]/u \in U\} \in N^f_{[w]}$,
2. For every formula $\#\varphi \in \Sigma$, if $\mathcal{U} \in N^f_{[w]}$ and $(\forall [u] \in \mathcal{U})(\mathfrak{M}, u \models \varphi)$, then $\mathfrak{M}, w \models \#\varphi$,
3. $V^f(p) = \{[w] \,/\, \mathfrak{M}, w \models p\}$, for all proposition letter p in Σ.

Condition (1) requires that for every neighbourhood of w there is a corresponding neighbourhood of classes of equivalences for the class of equivalence of w (i.e. $[w]$) in the filtration. Condition (2) settles, among classes of equivalences, the satisfaction definition regarding a world and its neighbourhoods.

We use U for the neighbourhoods in the original model $\mathfrak{M}$, and $\mathcal{U}$ for the neighbourhoods of $[w]$ in the filtration $\mathfrak{M}^f$.

Theorem 1 (Filtration Theorem for the neighbourhood case.)**.** *Consider a unary modality '#'. Let $\mathfrak{M}^f$ be a filtration of $\mathfrak{M}$ through a subformula closed set Σ. Then for all φ in Σ and all w in $\mathfrak{M}$, $\mathfrak{M}, w \models \varphi$ iff $\mathfrak{M}^f, [w] \models \varphi$. That is, filtration preserves satisfiability.*

Proof. We show that $\mathfrak{M}, w \models \varphi$ iff $\mathfrak{M}^f, [w] \models \varphi$. As Σ is subformula closed, we use induction on the structure of φ. We focus on the case $\varphi = \#\gamma$. Assume that $\#\gamma \in \Sigma$, and that $\mathfrak{M}, w \models \#\gamma$. If $\mathfrak{M}, w \models \#\gamma$ then there is a neighbourhood U such that $U \in N_w$ and $(\forall u \in U)(\mathfrak{M}, u \models \gamma)$, that is, for every world in that neighbourhood, γ holds. Thus, by application of the induction hypothesis, for each of those u we have that $\mathfrak{M}^f, [u] \models \gamma$. By condition (1) above, $\{[u]/u \in U\} \in N^f_{[w]}$. Hence $\mathfrak{M}^f, [w] \models \#\gamma$.

Conversely we have to prove that if $\mathfrak{M}^f, [w] \models \varphi$ then $\mathfrak{M}, w \models \varphi$.

Assume that $\varphi = \#\gamma$ and $\mathfrak{M}^f, [w] \models \#\gamma$. By truth definition, there exists $\mathcal{U}$ neighbourhood of $[w]$ such that $(\forall [u] \in \mathcal{U})(\mathfrak{M}^f, [u] \models \gamma)$. Then by inductive hypothesis $(\forall [u] \in \mathcal{U})(\mathfrak{M}, u \models \gamma)$. Then by condition (2) $\mathfrak{M}, w \models \#\gamma$. □

Note that clauses (1) and (2) above are devised to make the neighbourhood case of the induction step straightforward.

Existence of a filtration.

Notation. $[U] = \{[u]/u \in U\}$ i.e. $[U]$ is a set of classes of equivalences. Define $N^s_{[w]}$ as follows: $[U] \in N^s_{[w]}$ iff $(\exists\, w \equiv_\Sigma w'/U \in N_{w'})$. That is, $[U]$ is a neighbourhood of $[w]$ if there exists a neighbourhood U in the original model reachable through a world w' which is equivalent to w (under $\equiv_\Sigma$). This definition leads us to the smallest filtration.

Lemma 1 (See Lemma 2.40 in [3])**.** *Let $\mathfrak{M}$ be any model, Σ any subformula closed set of formulas, W_Σ the set of equivalence classes of W induced by $\equiv_\Sigma$, V^f the standard valuation on W_Σ. Then $\langle W_\Sigma, N^s_{[w]}, V^f\rangle$ is a filtration of $\mathfrak{M}$ through Σ.*

Proof. It suffices to show that $N^s_{[w]}$ fulfills clauses (1) and (2) in Definition 4. Note that it satisfies (1) by definition. It remains to check that $N^s_{[w]}$ fulfills (2).

Let $\#\varphi \in \Sigma$, we have to prove that $(\forall \mathcal{U} \in N^s_{[w]})$ $(\forall [u] \in \mathcal{U})(\mathfrak{M}, u \models \varphi) \rightarrow (\mathfrak{M}, w \models \#\varphi)$. We know that $\mathcal{U} = [U]$ for some $U \in N_{w'}$ such that $w \equiv_\Sigma w'$. Recall that $(\forall [u] \in \mathcal{U})(\mathfrak{M}, u \models \varphi)$ means that $(\forall u \in U)(\mathfrak{M}, u \models \varphi)$. By truth definition $\mathfrak{M}, w' \models \#\varphi$, then because $w \equiv_\Sigma w'$ we get $\mathfrak{M}, w \models \#\varphi$. □

Theorem 2 (Finite Model Property via Filtrations)**.** *Assume that φ is satisfiable in a model $\mathfrak{M}$ as in Definition 3; take any filtration $\mathfrak{M}^f$ through the set of subformulas of φ. That φ is satisfiable in $\mathfrak{M}^f$ is immediate from the Filtration Theorem for the neighbourhood case.*

Being $\equiv_\Sigma$ an equivalence relation, and using Theorem 1 it's easy to check that, a model $\mathfrak{M}$ and any filtration $\mathfrak{M}^f$ are equivalent modulus φ. This result is useful to understand why the original properties of the frames in the models are preserved. This results are provided in [Chellas] for the preservation of frames clases through filtrations.

Example 1 (uni-agent mono-modal system)**.** A simple system can be defined with structure as in Definition 3, where we can write and test situations like the one following:

Bus stop scenario ([13], revisited). Suppose that agent y is at the bus stop. We can test whether y raises his hand and stops the bus by testing the validity of the formula: $\mathrm{Does}_y(StopBus)$. This simple

kind of systems are proved decidable via FMP through Definition 4, Theorem 1 and Lemma 1 in this Section. They are powerful enough to monitor a single agent's behaviour.

Note that $\mathrm{Does}_y(StopBus)$ holds in a world w in a model $\mathfrak{M}$, that is, $\mathfrak{M}, w \models \mathrm{Does}_y(StopBus)$ iff $(\exists\, U \in N_{y_w})$ such that $(\forall\, u \in U)\,(\mathfrak{M}, u \models StopBus)$.

3 Extension to the multi-agent multi-modal case

Recall that the original base structure discussed in [32] is a multi-relational frame of the form:

$$\mathfrak{F} = \langle A, W, \{B_i\}_{i\in A}, \{G_i\}_{i\in A}, \{I_i\}_{i\in A}, \{D_i\}_{i\in A}\rangle$$

where:

- A is a finite set of agents;
- W is a set of situations, or points, or possible worlds;
- $\{B_i\}_{i\in A}$ is a set of accessibility relations wrt Bel, which are transitive, euclidean and serial;
- $\{G_i\}_{i\in A}$ is a set of accessibility relations wrt Goal, (standard K_n semantics);
- $\{I_i\}_{i\in A}$ is a set of accessibility relations wrt Int, which are serial; and
- $\{D_i\}_{i\in A}$ is a family of sets of accessibility relations D_i wrt Does, which are pointwise closed under intersection, reflexive and serial [17].

This original structure contains the well-known normal operators Bel, Goal, and Int. They have a necessity semantics, plus characterizing axioms (see for example [19, 9]). These operators are the ones we aim to arbitrarily combine with the non-normal Does.

Note that the necessity semantics for the Kripke case can be written using neighbourhood semantics in the following way (see [6] Theorem 7.9 for more detail):

$\mathfrak{M}^K, w \models \varphi$ iff $(\forall v \,/\, wRv)(\mathfrak{M}^K, v \models \varphi) \Longleftrightarrow \mathfrak{M}^N, w \models \varphi$ iff $(\forall\, v_k \in N_w)\,(\forall u \in v_k)(\mathfrak{M}^N, u \models \varphi)$

where $\mathfrak{M}^K$ is a Kripke model, and $\mathfrak{M}^N$ is a neighbourhood model.

The intuition behind this definition is that each world v accessible from w in $\mathfrak{M}^K$ is a neighbourhood of w in $\mathfrak{M}^N$. Standard models can be paired one-to-one with neighbourhood models in such a way that paired models are pointwise equivalent [6].

So we can think of having a $\{N_{i_w}\}$ for each normal modality, as we do for the Does modality.

Now let us consider a multi-modal system with structure $\langle W, \{N_{1_w}\}, ..., \{N_{m_w}\}\rangle$ and let us assume that we have one agent. It is straightfoward to extend the application of Theorem 1 (Section 2) to this structure. Asume a basic modal language with modalities $\#_1, ..., \#_m$, each with a neighbourhood semantics. Also, consider a set Σ closed for subformulas that satisfies: (i) if $\varphi \vee \varphi' \in \Sigma$ then $\varphi \in \Sigma$ and $\varphi' \in \Sigma$; (ii) if $\neg\varphi \in \Sigma$, then $\varphi \in \Sigma$; and (iii) if $\#_i\,\varphi \in \Sigma$, then $\varphi \in \Sigma$ for every $\#_i$.

Definition 5 (Extends Definition 4)**.** Let $\mathfrak{M} = \langle W, \{N_{1_w}\}, ..., \{N_{m_w}\}, V\rangle$ be a model, Σ a subformula closed set, $\equiv_\Sigma$ an equivalence relation. Let $\mathfrak{M}^f = \langle W^f, \{N_{1_w}\}^f, ..., \{N_{m_w}\}^f, V^f\rangle$ such that $W^f = W_\Sigma$ and:

1. If $U \in N_{i_w}$ then $\{[u]/u \in U\} \in N^f_{i_{[w]}}$; and
2. For every formula $\#_i\,\varphi \in \Sigma$, if $\mathcal{U} \in N^f_{i_{[w]}}$ and $(\forall [u] \in \mathcal{U})(\mathfrak{M}, u \models \varphi)$, then $\mathfrak{M}, w \models \#_i\,\varphi$.
3. $V^f(p) = \{[w] \,/\, \mathfrak{M}, w \models p\}$, for all proposition letter p in Σ.

It is easy to check that if Σ is a subformula closed set of formulas, then $\mathfrak{M}^f$ is a filtration of $\mathfrak{M}$ through Σ. That is, for all φ in Σ and all w in $\mathfrak{M}$, $\mathfrak{M}, w \models \varphi$ iff $\mathfrak{M}^f, [w] \models \varphi$. Proof is done by repeated application of Theorem 1 (Section 2). Clearly, it suffices to prove the result for a single '$\#_i$' as all modalities have a neighbourhood semantics. It is worth mentioning that authors in [10], for example, proceed with the direct repeated application of the notion of filtration for proving the FMP of their (normal) multi-modal system.

Example 2 (uni-agent multi-modal system)**.** A simple system can be defined according to Definition 5, where we can depict scenarios and test situations like the one following:

Bus stop example (revisited). Agent x is at the bus stop having the goal to stop the bus: $\mathrm{Goal}_x(\mathrm{Does}_x(StopBus))$.

Note that $\mathrm{Goal}_x(\mathrm{Does}_x(StopBus))$ holds in a world w in a model $\mathfrak{M}$, that is, $\mathfrak{M}, w \models (\mathrm{Goal}_x\,\mathrm{Does}_x(StopBus))$ iff $(\exists\, U \in N_{x_w})$ such that $(\forall\, u \in U)(\mathfrak{M}, u \models \mathrm{Does}_x(StopBus))$, and $(\forall\, u \in U)(\mathfrak{M}, u \models \mathrm{Does}_x(StopBus))$ iff $(\exists\, U' \in N_{y_u})$ such that $(\forall\, u' \in U')(\mathfrak{M}, u' \models StopBus)$.

Further extension: multi-agent case

Extending the system to many agents will not add anything substantially new to Definition 5. A multi-agent system is a special case of the multi-modal case; the structure is merely extended with the inclusion of new modalities. For example, include Bel_i, Goal_i, and Int_i, for each agent i and a Does_i for each agent i. Thus, for every agent, include its corresponding modalities, each of which brings in its own semantics.

Example 3 (multi-agent multi-modal system)**.** A multi-agent multi-modal system for the bus stop scenario is, for example:

Bus stop example (re-revisited). The formula $\mathrm{Bel}_x(\mathrm{Does}_y(StopBus)))$ stands for 'agent x believes that agent y will stop the bus', meaning that he thinks he will not have to raise the hand himself. This formula holds in a world w in a model $\mathfrak{M}$, that is, $\mathfrak{M}, w \models \mathrm{Bel}_x\,\mathrm{Does}_y(StopBus)$ iff $(\exists\, U \in N_{x_w})$ such that $(\forall\, u \in U)(\mathfrak{M}, u \models \mathrm{Does}_y(StopBus))$, and $(\forall\, u \in U)(\mathfrak{M}, u \models \mathrm{Does}_y(StopBus))$ iff $(\exists\, U' \in N_{y_u})$ such that $(\forall\, u' \in U')(\mathfrak{M}, u' \models StopBus)$.

Another example.

Bus stop example (persuasion). $\mathrm{Does}_x(\mathrm{Goal}_y(StopBus))$ can be seen as a form of persuasion, meaning that 'agent x makes agent y stop the bus'. $\mathrm{Does}_x(\mathrm{Goal}_y(StopBus)))$ holds in a world w in a model $\mathfrak{M}$, that is, $\mathfrak{M}, w \models \mathrm{Does}_x\,\mathrm{Goal}_y(StopBus)$ iff $(\exists\, U \in N_{x_w})$ such that $(\forall\, u \in U)(\mathfrak{M}, u \models \mathrm{Goal}_y(StopBus))$, and $(\forall\, u \in U)(\mathfrak{M}, u \models \mathrm{Goal}_y(StopBus))$ iff $(\exists\, U' \in N_{y_u})$ such that $(\forall\, u' \in U')(\mathfrak{M}, u' \models StopBus)$.

Recall that we could not write and test *wff* with modalities within the scope of a Does in [32] and [31]. $\mathrm{Does}_i(\mathrm{Goal}_j\,\mathscr{A})$ is a formula in which the normal modality appears within the scope of a (non-normal) Does.

4 Combination of Mental States and Actions

Up to now, we described MAS under a single point of view: in this situation an agent believes this way, and acts that way. We are now interested in describing systems in which two points of view coexist: a cognitive one, and a behavioural one. These differ from the former ones on the ontology adopted.

We already referred in the Introduction that it is common to combine agent's behaviour with time. As a further example, a combination between a basic temporal and a simple deontic logic for MAS has been recently depicted in [33]. That combination puts together two normal modal logics: a temporal one and a deontic one. In the resultant system it is possible to write and test the validity of formulas with arbitrarily interleaved deontic and tense modalities. There are two structures (W, R) and $(T, <)$ which are respectively the underlying ontologies where a deontic point of view and a temporal point of view are interpreted (both are Kripke models). (W, R) represents a multigraph over situations, $(T, <)$ represents a valid time line. Next, it is built an ontology $W \times T$ of pairs (*situation*, *point in time*) representing the intuition "this situation, at this time". We note that such combination can be seen as a special case of the structure that we outline next. This outline (which is more general) allows combinations of non-normal operators having neighbourhood semantics.

For simplifying our presentation, we work again with the less possible number of modalities (say just two). We choose a normal, cognitive modality (let us say Bel, for beliefs), and a non-normal behavioural one (let us say Does, for agency).

Proposition 1. *If* $\langle W_B, \{N_B\}_{b \in W_B} \rangle$ *and* $\langle W_D, \{N_D\}_{d \in W_D} \rangle$ *are neighbourhood frames, then:*

$\mathfrak{C} = \langle W_B \times W_D, \{N_B\}_{(b,d) \in W_B \times W_D}, \{N_D\}_{(b,d) \in W_B \times W_D} \rangle$ *is a combined frame, where:*

- $W_B \times W_D$ *is a set of pairs of situations;*
- $S \in N_{B_{(b,d)}}$ *iff* $S = m \times \{d\}$, $m \in N_{B_b}$*; and*
- $T \in N_{D_{(b,d)}}$ *iff* $T = \{b\} \times n$, $n \in N_{D_d}$.

At a point (w_B, w_D) we have a pair of situations which are, respectively, environmental support for an internal configuration and for an external one. According to both dimensions, we test the validity of *wffs*: beliefs are tested on w_B and throughout the neigbourhoods of w_B provided by dimension S. The S dimension keeps untouched the behavioral dimension bound to w_B i.e. w_D is the second component on the neighbourhood S of w_B. (respectively for w_d and T).

In its turn, a combined model is a strucure $\langle \mathfrak{C}, V \rangle$ where V is a valuation function defined as expected. It is plain to see that this structure is an instance of Definition 5. That means there exists a flitration for a model based on this structure.

A MAS with structure as in Proposition 1 is said to be two-dimensional in the sense given by Finger and Gabbay in [14]: the alphabet of the system's language contains two disjoint sets of operators, and formulas are evaluated at a two-dimensional assignment of points that come from the prime frames' sets of situations. Moreover, in this "Beliefs $\times$ Actions" outline, there is no strong interaction among the logic of beliefs and the logic of agency as we define no interaction axioms among both special purpose logics. Our Proposition 1 much resembles the definition of full join given in [14] (Def 6.1) (two-dimensional plane).

Example 4 (Uni-agent combined system)**. Agent's beliefs and actions.** According to Proposition 1, we can define a system where to write and test formulas like e.g. $\mathrm{Bel}_x(\mathrm{Does}_x(\mathrm{Bel}_x\, \mathscr{A}))$. This formula is meant to stand for "agent x believes that s/he does what s/he believes" which can be seen as a kind of "positive introspection" regarding agency. This formula is not to be understood as an axiom bridging agency and beliefs; nonetheless it may be interesting to test its validity in certain circumstances: one may indeed believe that one is doing what meant to (expected correspondence between behaviour and belief), while one may believe one is doing something completely different to what one is effectively doing (e.g. poisoning a plant instead of watering it; or some other forms of erratic behaviour). Moreover, there are occasions where one performs an action which one does not believes in (e.g. obeying immoral orders).

For testing such formula, one possible movement along the multigraph is:

$\mathfrak{M}, (w_B, w_D) \models \mathrm{Bel}_x(\mathrm{Does}_x(\mathrm{Bel}_x\, \mathscr{A}))$ iff $(\exists\, U \in N_{B_{(w_B, w_D)}})$ such that $(\forall\, (u, w_D) \in U)$ $(\mathfrak{M}, (u, w_D) \models \mathrm{Does}_x(\mathrm{Bel}_x\, \mathscr{A})$. In its turn, $(\mathfrak{M}, (u, w_D) \models \mathrm{Does}_x(\mathrm{Bel}_x\, \mathscr{A})$ iff $(\exists\, \vee \in N_{D_{(u, w_D)}})$ such that $(\forall\, (u, v) \in \vee)$ $(\mathfrak{M}, (u, v) \models \mathrm{Bel}_x\, \mathscr{A})$. Finally, $(\mathfrak{M}, (u, v) \models \mathrm{Bel}_x\, \mathscr{A})$ iff $(\exists\, U' \in N_{B_{(u,v)}})$ such that $(\forall\, (u', v) \in U')$ $(\mathfrak{M}, (u', v) \models \mathscr{A})$.

In connection with our Example 4, it is worth mentioning that J. Broersen defines and explains in [5] a particular logics for doing something (un)knowingly. In that work (Section 3) the author explicitly defines some constraints for the interaction between knowledge and action, namely (1) an axiom that reflects that agents can not knowingly do more than what is affected by the choices they have, and (2) an axiom establishing that if agents knowingly see to it that a condition holds in the next state, in that same state agents will recall that such condition holds. The frames used are two-dimensional, with a dimension of histories (linear timelines) and a dimension of states agents can be in. Behaviours of agents can be interpreted as trajectories going from the past to the future along the dimension of states, and jumping from sets of histories to subsets of histories (choices) along the dimension of histories.

5 Conclusions

The idea of combining special purpose logics for building "on demand" MAS is promising. This engineering approach is, in this paper, balanced with the aim to handle decidable logics, which is a basis for the implementation and launching of correct systems. We believe that the decidability issue should be a prerequisite to be taken into account during the design phase of MAS.

Within the MAS community the neighbourhood notation is, possibly, most widely used, well-understood, and well-recognized than the selection function notation. We gave a "neighbourhood outline" to decidability via filtration for a particular kind of models, namely neighbourhood models. These models are suitable for capturing the semantics of some non-normal operators found in the MAS literature (such as agency, or ability, among others) and, of course, also the semantics of normal modal operators as most MAS use.

We also offered technical details for combining logics which can be used as a basis for modeling multi-agent systems. The logics resulting from different possible combinations lead to interesting levels of expressiveness of the systems, by allowing different types of complex formulas. The combinations outlined in this paper are, given the logical tools presented in Section 2, decidable. There are for sure several other possible combinations that can be performed. For exam-

ple, Proposition 1 can be extended to capture more cognitive aspects such as e.g. goals, or intentions. In that case, the cognitive dimension (In Proposition 1, characterized by S) is to be extended with the inclusion of normal operators. Moreover, within our neighbourhood outline and on top of the uni-agent modalities, collective modalities such as mutual intention, collective intention; also elaborated concepts such as trust or collective trust can also be defined.

We can push the combination strategy even further, by proposing the combination of modules which are in its turn combinations of special purpose logics, in a kind of multiple level combination. This strategy has to be carefully studied, and is matter of our future research.

REFERENCES

[1] C. Areces, C. Monz, H. de Nivelle, and M. de Rijke, 'The guarded fragment: Ins and outs', in *JFAK. Essays Dedicated to Johan van Benthem on the Occasion of his 50th Birthday*, eds., J. Gerbrandy, M. Marx, M. de Rijke, and Y. Venema, Vossiuspers, AUP, (1999).

[2] Patrick Blackburn, Johan F. A. K. van Benthem, and Frank Wolter, *Handbook of Modal Logic, Volume 3 (Studies in Logic and Practical Reasoning)*, Elsevier Science Inc., New York, NY, USA, 2006.

[3] Patrick Blackburn, Maarten de Rijke, and Yde Venema, *Modal Logic*, volume 53 of *Cambridge Tracts in Theoretical Computer Science*, Cambridge University Press, Cambridge, 2001.

[4] Jan Broersen, 'Relativized action complement for dynamic logics', in *Advances in Modal Logic*, pp. 51–70, (2002).

[5] Jan Broersen, 'A complete stit logic for knowledge and action, and some of its applications', in *DALT*, pp. 47–59, (2008).

[6] Brian Chellas, *Modal Logic: An Introduction*, Cambridge University Press, 1980.

[7] Philip R. Cohen and Hector J. Levesque, 'Intention is choice with commitment', *Artif. Intell.*, **42**(2-3), 213–261, (March 1990).

[8] J. W. de Bakker, Willem P. de Roever, and Grzegorz Rozenberg, eds. *Linear Time, Branching Time and Partial Order in Logics and Models for Concurrency, School/Workshop, Noordwijkerhout, The Netherlands, May 30 - June 3, 1988, Proceedings*, volume 354 of *Lecture Notes in Computer Science*. Springer, 1989.

[9] Barbara Dunin-Keplicz and Rineke Verbrugge, 'Collective intentions.', *Fundam. Inform.*, 271–295, (2002).

[10] Marcin Dziubiński, Rineke Verbrugge, and Barbara Dunin-Kęplicz, 'Complexity issues in multiagent logics', *Fundam. Inf.*, **75**(1-4), 239–262, (January 2007).

[11] Dag Elgesem, 'The modal logic of agency', *Nordic Journal of Philosophical Logic*, **2**, 1–46, (1997).

[12] Rogerio Fajardo and Marcelo Finger, 'Non-normal modalisation.', in *Advances in Modal Logic'02*, pp. 83–96, (2002).

[13] R. Falcone and C. Castelfranchi, *Trust and Deception in Virtual Societies*, chapter Social Trust: A Cognitive Approach, 55–90, Kluwer Academic Publishers, 2001.

[14] Marcelo Finger and Dov Gabbay, 'Combining temporal logic systems', *Notre Dame Journal of Formal Logic*, **37**, (1996).

[15] Massimo Franceschet, Angelo Montanari, and Maarten De Rijke, 'Model checking for combined logics with an application to mobile systems', *Automated Software Eng.*, **11**, 289–321, (June 2004).

[16] Dov Gabbay, *Fibring Logics*, volume 38 of *Oxford Logic Guides*, Oxford Univesity Press, 1998.

[17] Guido Governatori and Antonino Rotolo, 'On the Axiomatization of Elgesem's Logic of Agency and Ability', *Journal of Philosophical Logic*, **34**(4), 403–431, (2005).

[18] Guido Governatori and Antonino Rotolo, 'Norm compliance in business process modeling', in *Semantic Web Rules*, eds., Mike Dean, John Hall, Antonino Rotolo, and Said Tabet, volume 6403 of *Lecture Notes in Computer Science*, 194–209, Springer Berlin / Heidelberg, (2010).

[19] J. Halpern and Y. Moses, 'A guide to completeness and complexity for modal logics of knowledge and belief', *Artificial Intelligence*, **54**, 311–379, (1992).

[20] Bengt Hansson and Peter Gärdenfors, 'A guide to intensional semantics. Modality, Morality and Other Problems of Sense and Nonsense. Essays Dedicated to Sören Halldén', (1973).

[21] Philip R. Cohen Hector J. Levesque and Jose H. T. Nunes, 'On acting together', Technical Report 485, AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, (May 1990).

[22] H. Herrestad and C. Krogh, *Deontic Logic Relativised to Bearers and Counterparties*, 453–522, J. Bing and O. Torvund, 1995.

[23] A.J.I. Jones and M. Sergot, 'A logical framework. in open agent societies: Normative specifications in multi-agent systems', (2007).

[24] Dexter Kozen, 'Results on the propositional mu-calculus', *Theor. Comput. Sci.*, **27**, 333–354, (1983).

[25] J. J. Ch Meyer, 'A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic', *Notre Dame Journal of Formal Logic*, **29**(1), 109–136, (1987).

[26] Jeremy Pitt, 'A www interface to a theorem prover for modal logic', in *Department of Computer Science, University of York*, pp. 83–90, (1996).

[27] ANAND S. RAO and MICHAEL P. GEORGEFF, 'Decision procedures for bdi logics', *Journal of Logic and Computation*, **8**(3), 293–343, (1998).

[28] Antonino Rotolo, Guido Boella, Guido Governatori, Joris Hulstijn, Regis Riveret, and Leendert van der Torre, 'Time and defeasibility in FIPA ACL semantics', in *Proceedings of WLIAMAS 2008*. IEEE, (2008).

[29] Klaus Schild, 'On the relationship between bdi logics and standard logics of concurrency', *Autonomous Agents and Multi-Agent Systems*, **3**(3), 259–283, (September 2000).

[30] Peter K. Schotch, 'Paraconsistent logic: The view from the right', in *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume Two: Symposia and Invited Papers, pp. 421–429. The University of Chicago Press, (1992).

[31] Clara Smith, Agustin Ambrossio, Leandro Mendoza, and Antonino Rotolo, 'Combinations of normal and non-normal modal logics for modeling collective trust in normative mas', *AICOL XXV IVR, Forthcoming LNAI*, (2011).

[32] Clara Smith and Antonino Rotolo, 'Collective trust and normative agents', *Logic Journal of IGPL*, **18**(1), 195–213, (2010).

[33] Clara Smith, Antonino Rotolo, and Giovanni Sartor, 'Representations of time within normative MAS', *Frontiers in Artificial Intelligence and Applications*, **223**, 107–116, (2010).

The Effects of Social Ties on Coordination: Conceptual Foundations for an Empirical Analysis

Giuseppe Attanasi[1] and **Astrid Hopfensitz**[1] and **Emiliano Lorini**[2] and **Frédéric Moisan** [2]

Abstract. In this paper, we are investigating the influence that social ties can have on behavior. After first defining the concept of social ties that we consider, we propose a coordination game with outside option, which allows us to study the impact of such ties on social preferences. We provide a detailed game theoretic analysis of this game while considering various types of players: i.e. self-interest maximising, inequity averse, and fair agents. Moreover, in addition to these approaches that require strategic reasoning in order to reach some equilibrium, we also present an alternative hypothesis that relies on the concept of team reasoning. Finally, we show that an experiment could provide insight into which of these approaches is the most realistic.

1 Introduction

In classical economic theories, most models assume that agents are self-interested and maximize their own material payoffs. However, important experimental evidence from economics and psychology have shown some persistent deviation from such self-interested behavior in many particular situations. These results suggest the need to incorporate social preferences into game theoretical models. Such preferences describe the fact that a given player not only considers his own material payoffs but also those of other players [27]. The various social norms created by the cultural environment in which human beings live give us some idea of how such experimental data could be interpreted: fairness, inequity aversion, reciprocity and social welfare maximization all represent concepts that everybody is familiar with, and which have been shown to play an important role in interactive decision making (e.g. see [16, 11, 28]).

In fact, various simple economic games, such as the trust game [4] and the ultimatum game [23], have been extensively studied in the past years because they illustrate well the weakness of classical game theory and its assumption of individualistic rationality. Moreover, given the little complexity carried out in such games, the bounded rationality argument [19] does not seem sufficient to justify the observed behaviors. Social preferences appear as a more realistic option because it allows to explain the resulting behaviors while still considering rational agents.

However, although many economic experimental studies (e.g. [4, 23]) have shown that people genuinely exhibit other-regarding preferences when interacting with perfect strangers, one may wonder to what extent the existence of some social ties between individuals may influence behavior. Indeed the dynamic aspect of social preferences seems closely related to that of social ties: one may cooperate more with a friend than with a stranger, and doing so may eventually enforce the level of friendship. Yet, in spite of their obvious relevance to the study of human behavior, very little is known about the nature of social ties and their actual impact on social interactions.

Our attempt, through this paper, is to study the possible effects that positive social ties can have on human cooperation and coordination. Our main hypothesis is that such relationships can directly influence the social preferences of the players: an agent may choose to be fair conditionally to the relative closeness with his opponent(s). In order to investigate this theory, we propose a theoretical analysis of a specific two player game, which creates an ideal context for the study of social ties and social preferences.

The rest of the article is organized as follows. Section 2 defines the concept of a social tie that we consider. In section 3, we propose a game that allows to measure the behavioral effects of social ties. We then provide in Section 4 a game theoretical analysis of this game by considering only self-interested agents. Then in Section 5, we perform a similar analysis by considering other-regarding agents according to theories of social preferences. Finally, in Section 6, we propose an alternative interpretation of the same game by considering agents as team-directed reasoners.

2 A basic theory of social ties

As previously mentioned, there exists no formal definition of a social tie in the literature, and this is why, given the vagueness and the ambiguity that the term may suggest, we first have to clarify the concept that we consider.

First, we choose to restrict our study only to those ties that can be judged to be positive: examples of those include relationships between close friends, married couples, family relative, class mates, etc... In contrast, negative ties may include relationships between people with different tastes, from different political orientations, with different religious beliefs, etc...

It seems reasonable to compare this concept of a social tie with social identity theory from social psychology [34]. In fact, the existence of a bond between two individuals seems likely to make them identify themselves to the same social group, whatever such a group might be. However, whether belonging to the same group actually implies the existence of some social tie remains unclear. To illustrate this point, let us consider the Minimal Group Paradigm (MGP) [34], which corresponds to an experimental methodology from social psychology that investigates the minimal conditions required for discrimination to occur between groups. In fact, experiments using this approach [35] have revealed that arbitrary and virtually meaningless distinctions between groups (e.g. the colour of their shirts) can trigger a tendency to cooperate more with one's own group than with others. One meaningful interpretation from such results is that prejudice can indeed have some non negligible influence on social behavior.

[1] Toulouse School of Economics (TSE)
[2] Université de Toulouse, CNRS, Institut de Recherche en Informatique de Toulouse (IRIT)

This brings us to focus on the intrinsic foundations of social ties and the possible reasons for their emergence. Following the previous studies based on the MGP, it is reasonable to state that social ties rely, at least to some extent, on sharing some common social features. One can then distinguish the following dimensions of proximity:

- Similarity of features: i.e. sharing the same social features (belonging to the same political party, having the same religious orientation, etc ...) One should note that this categorization may include any form of prejudice.
- Importance of features: i.e. the degree of importance people give to particular social features (the importance given to belonging to some political party, the degree of faith in some particular religion, etc ...)

One should note that correlation across such social features can sometimes suffice to imply the same behavior: for instance, activists from the same political party may share some fairness properties. Moreover, experimental studies in economics [21, 12] suggest that such social proximity between interacting individuals may induce group identity and therefore directly affect social preferences and norm enforcement.

As a concrete example to illustrate the above theory of common social features, one may consider the approach by online dating systems (as currently flourishing on the internet). In fact, those systems, which are clearly meant to build social ties between individuals (assuming an affective tie as a special case of a social tie), are clearly based on the matching of both the similarity and the importance of features. However, while one cannot deny the effectiveness of such systems [24], it is doubtful to assume that such criteria are sufficient to fully define social ties [17].

The previous example suggests that social ties require some additional sharing of information, which help identify particular behavioral patterns. In fact, human beings are learning agents that genuinely infer judgements from experience. One may then assume that social ties also rely on some experience-based proximity, which involves actual interactions between individuals. For example, eliciting some altruistic behavior in some interactive situation may be likely to contribute to the creation of a social bond with other individuals. One should note that the main difference here with the other dimensions of proximity described above lies in the necessity to observe the others' behaviors during past interactions.

The last issue that we wish to address here concerns the bilateral and symmetric aspect of a social tie. Indeed, any unilateral bond should be simply understood as some belief about the existence of a social tie: as an example, although Alice can see the same TV-show host every day and knows that they both share some common social features, there cannot be any social tie as long as the TV host does not know her.

As a consequence, this leads to the following hypothesis:

Statement 2.1 *a social tie (to a certain degree k) exists between two individuals* ***if and only if*** *both individuals commonly believe that the tie exists.*

3 The social tie game

Having previously analysed the main characteristics of a social tie, we now propose a game that seems best suited to study its behavioral effects.

The corresponding Social Tie (ST) game, which is shown in Figure 1, is a two player game that can be described as follows: during the first stage of the game, only Alice has to choose between either playing In or Out. In the latter case, the game ends with Alice earning \$20 and Bob earning \$10. In the former case (i.e. In), both players enter the second stage of the game that corresponds to a basic coordination game. If both coordinate on the (C_a, C_b) solution, then Alice and Bob get \$35 and \$5, respectively. Similarly, if both coordinate on the (D_a, D_b) solution, then they get \$15 (Alice) and \$35 (Bob). In any other case, both players win nothing (\$0).

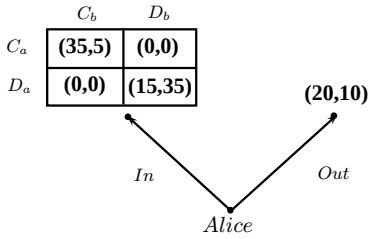


Figure 1. Social Tie game

One may note that our ST game corresponds to a variant of the Battle of the Sexes (BoS) game with outside option (see [15]). Indeed, the only difference lies on the symmetrical property within the coordination subgame that we voluntarily removed here: unlike in the BoS game, the lowest payoff is different in the two coordination outcomes (\$5 $\neq$ \$15). The main motivation to introduce this type of asymmetry is to create some incentives for the players to favour the group as a whole (in fact, neither social preferences, not team reasoning would affect behavior in a BoS-like subgame).

One may also notice the similarity with the Dalek game presented in [5]. The only difference with our ST game is that in the Dalek game one solution of the coordination subgame ensures perfect equity. Indeed, as in our case, the Dalek game also introduces some dilemma between maximizing one's self-interest and playing the fairest outcome. However, unlike in our ST game, it does not introduce any dilemma between satisfying self-interest and maximizing the social welfare (i.e. the combined payoffs of every player). Although this game would be interesting to investigate, it may also make it more difficult to observe the actual effects of social ties on behavior: as a consequence of this missing dilemma, the Dalek game offers less incentive to play the fairest solution, which may eventually lead to a higher rate of miscoordination, with and without the presence of such ties. On the other hand, the signal of perfect equity in the Dalek game may also appear so strong that it could reinforce the stability of coordinating on the corresponding solution, even when no ties are involved.

4 Game theoretic analysis

Through this section, we wish to provide a full theoretical analysis of the above ST game that is exclusively based on classical game theory (i.e. assuming agents are self-interested maximizers). In order to do so, we will define the sets of Nash equilibria, subgame perfect equilibria, and forward induction solutions.

4.1 Nash equilibria

First consider the coordination subgame alone (i.e. the second stage of the full ST game). Such a game has three different Nash equilibria – two asymmetric ones in pure strategies, (C_a, C_b) and (D_a, D_b), and one in mixed strategies in which Alice plays C_a with probability $7/8$ and earns an expected payoff of \$10.5, while Bob plays C_b with probability $3/10$ and earns an expected payoff of \$4.375.

Let us now consider the full ST game, which consists of the previous coordination game extended with some outside option (at the first stage of the game). The corresponding game in normal form is represented in Figure 2.

	C_b	D_b
(In, C_a)	**(35,5)**	**(0,0)**
(In, D_a)	**(0,0)**	**(15,35)**
(Out, C_a)	**(20,10)**	**(20,10)**
(Out, D_a)	**(20,10)**	**(20,10)**

Figure 2. Social Tie game in normal form

This game contains three Nash equilibria in pure strategies, which are the following:

$$(In, C_a, C_b), (Out, C_a, D_b), (Out, D_a, D_b)$$

These equilibria should simply be understood as follows: As long as Bob does play D_b in the coordination subgame, then Out remains the best option for Alice (no matter what Alice would have chosen between C_a and D_a in the subgame). In any other cases, the strategy (In, C_a) becomes the only rational move for Alice.

One should note that this set of solutions should be extended by a large number of Nash equilibria in mixed strategies: we voluntarily postpone the analysis of such solutions to the next section.

4.2 Subgame perfect equilibria

The subgame perfect equilibria, which can be computed through the backward induction method, represent a restriction on the previous set of Nash equilibria. In fact, this solution concept allows to rule out incredible solutions that may be predicted as Nash equilibria. In our game, (Out, C_a, D_b) represents such a solution. Indeed, although the prediction to play Out is perfectly rational for Alice, it here relies on the fact that she would not be rational if she had played In in the first place: given that Bob plays D_b in the coordination subgame, Alice's only rational move would be to play D_a instead of C_a (which corresponds to a Nash equilibrium in the subgame).

Moreover, one should note that the backward induction principle also discard every Nash equilibrium in mixed strategies. In fact, the optimal mixed strategy in the coordination subgame (see Section 4.1) is strictly dominated by the Out option.

As a consequence, the set of all subgame perfect Nash equilibria in pure strategies reduces to the following:

$$(In, C_a, C_b), (Out, D_a, D_b)$$

4.3 Forward induction

Similarly the forward induction principle restricts the previous set of subgame perfect Nash equilibria to keep only the most rational solutions, which resist the iteration of weak dominance. In the context of our ST game (see Figure 2), this leads to the following solution: first Alice's strategy (In, D_a) is weakly (and strictly) dominated by any strategy involving Out. Then Bob's strategy D_b becomes weakly dominated by D_a. Thus Alice's strategies (Out, C_a) and (Out, D_a) are both weakly (and strictly) dominated by (In, C_a, C_b). Therefore, the unique forward induction solution, which resist iterated weak dominance, is as follows:

$$(In, C_a, C_b)$$

Indeed it turns out that fully rational players should play this solution, which can be interpreted as follows: while playing In, Alice signals Bob that she intends to play C_a (if she intended to play D_a, she would have played Out in the first place). Therefore Bob's only rational move is then to play C_b. However, while this interpretation justify the existence of the above solution, it does not explain why the other backward induction solution is not rational. To continue the argument, let us then consider the solution (Out, D_a, D_b), which can be interpreted as follows: Alice plays Out because she expects Bob to play D_b in case she had played In. This chain of reasoning is clearly erroneous because Alice's conditional expectation does not match what she would really expect if she had *actually* chosen to perform In. Indeed, as shown before, if Alice performs In, Bob's only rational move is to play C_b, so no matter what Alice does during the first stage, she cannot expect anything else than Bob playing C_b. Consequently, her only rational move is to play (In, C_a), and Bob's best response is to play (C_b).

The interesting characteristics that this analysis brings about is that the validity of this forward induction argument is independent on Bob's preferences. This therefore suggests that such a game introduces some "first mover" advantage that the second player can not exploit, assuming that it is common knowledge among them that they both are self interested agents.

Many studies in the economic literature have shown support to this forward induction argument, see e.g. [8, 31, 14, 15, 36, 9, 10, 3].

Cooper et al. [14] investigate a coordination game with two Pareto-ranked equilibria and report that a payoff-relevant outside option changes play in the direction predicted by forward induction. Van Huyck et al. [36] report the success of forward induction in a setup in which the right to participate in a coordination game is auctioned off prior to play. Cachon and Camerer [10] investigate a setup in which subjects may pay a fee to participate in a coordination game with Pareto-ranked equilibria. They report that play is consistent with forward induction.

However, there is also contrary evidence. In [15], Cooper et al. obtain the forward induction solution when it coincides with a dominance argument but the same outcome is predicted when forward induction makes no prediction. Brandts and Holt [9] also show that the forward induction is only a good prediction if it coincides with a simple dominance argument. In [7], Brandts et al. find evidence against forward induction in an industrial organization game.

Other work have shown that the temporal factor of the game is relevant to the forward induction reasoning. In [15] and [25], the forward induction solution predicts well in the experiment based on the extensive form but does poorly when subjects are presented with the normal form game.

However, all these work consider games that are slightly different from our current version. One may then wonder whether the asymmetry introduced in our ST game does resist the game theoretical prediction.

5 Introducing social preferences

In this section, we reinterpret our ST game through the use of existing economic theories of social preferences. In fact, these models allows one to consider not only the self-interested motivations of the agents, but also their social motivations. In other words, a player's utility is not characterised by his own material payoffs, but also those of the other players. We choose to focus on the concepts of inequity aversion and fairness, which seem to be the most relevant to our current game. Other models of reciprocity and altruism do not appear to be suitable to such a coordination game: those models would indeed re-

quire agents to predict the opponent's move and behave in a way that would be indistinguishable from that of some self-interested agent.

5.1 Theory of inequity Aversion

In the models proposed by Fehr & Schmidt [16] and Bolton & Ockenfels [6], players are assumed to be intrinsically motivated to distribute payoffs in an equitable way: a player dislikes being either better off or worse off than another player. In other terms, utilities are calculated in such a way that equitable allocations of payoffs are preferred by all players.

Formally, consider two players i and j and let $x = \{x_i, x_j\}$ denote the vector of monetary payoffs. According to Fehr & Schmidt's model, the utility function of player i is given by:

$$U_i(x) = x_i - \alpha_i \cdot max\{x_j - x_i, 0\} - \beta_i \cdot max\{x_i - x_j, 0\}$$

where it is assumed that $i \neq j$, $\beta_i \leq \alpha_i$ and $0 \leq \beta_i < 1$.

The two parameters can be interpreted as follows: α_i parametrizes the distaste of person i for disadvantageous inequality while β_i parametrizes the distaste of person i for advantageous inequality. One should note that setting these parameters to zero defines some purely self-interested agent. The constraints imposed on the parameters are meant to ensure that players do not act altruistically, which is not the purpose of the model (i.e. if $\alpha_i < \beta_i$ then the model would assume i is altruist).

Clearly, applying such a model to our current ST game can literally transform its whole structure, depending on the values assigned to parameters α_i and β_i. Let us then perform a game theoretical analysis that involves such inequity aversion parameters.

The main observation that can be made is about the effects of Alice's preference ordering on her behavior. In fact, assuming that $\beta_{Alice} \leq \alpha_{Alice}$, then Alice will never play the strategy (In, D_a), no matter how inequity averse she is:

- if $\beta_{Alice} < 3/4$ and $\alpha_{Bob} < 1/6$, then Alice and Bob's preferences remain as if they were self-interested (i.e. the forward induction argument still holds). Thus Alice's only rational strategy is to play (In, C_a) while Bob will rationally play (C_b).
- if $\beta_{Alice} < 3/4$ and $\alpha_{Bob} \geq 1/6$, then Alice is always better off by playing (Out): the coordination subgame yields a unique Nash equilibrium (i.e. (D_a, D_b)), which is strictly dominated by playing (Out).
- if $\beta_{Alice} \geq 3/4$, then Alice is always better off by playing (Out): for any $\alpha_{Alice} \geq \beta_{Alice}$, any outcome from the coordination subgame is strictly dominated by playing (Out) (see Figure 3 for an example).

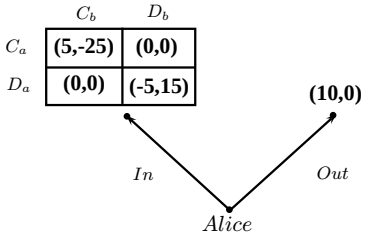


Figure 3. Transformed ST game with inequity averse players ($\alpha_{Alice} = \beta_{Alice} = \alpha_{Bob} = \beta_{Bob} = 1$)

The main result of this analysis is that the value of α_{Alice} and β_{Bob} are irrelevant to defining Alice and Bob's preferences. In other words, only Alice's distaste about advantageous inequality can affect her preference ordering in the current game. Similarly, only Bob's distaste about disadvantageous inequality can affect his preference ordering. One should also note that inequity aversion does not keep the "first mover" advantage mentioned in the previous section: Alice's first move does signal Bob not only about her low level of inequity aversion, but also about her expectation of Bob's low level of inequity aversion. That means that if she plays In, then the resulting outcome is entirely depending on Bob's level of inequity aversion (either (In, C_a, C_b) or (In, C_a, D_b) will be played).

The set of Nash Equilibria (NE) and Subgame Perfect Equilibria (SPE), in the context of the ST game played with inequity aversion, is summarized through the following table (note that forward induction is irrelevant in this case because the SPE always predicts a unique solution).

NE	SPE
(Out, C_a, C_b)	(Out, C_a, C_b) if $\alpha_{Bob} < 1/6$
(Out, C_a, D_b)	(Out, D_a, D_b) if $\beta_{Alice} < 3/4$
(Out, D_a, D_b)	(Out, C_a, D_b) if $\alpha_{Bob} \geq 1/6$ and $\beta_{Alice} \geq 3/4$
(Out, D_a, C_b)	

Table 1. Equilibrium solution concepts for inequity averse agent(s) ($\beta_{Alice} \geq 3/4$ **or** $\alpha_{Bob} \geq 1/6$)

5.2 Theory of fairness

Let us now consider another type of social preferences model that relies on the notion of fairness. In [11], Charness & Rabin propose a specific form of social preference they call *quasi-maximin* preferences. In their model, group payoff is computed by means of a social welfare function which is a *weighted* combination of Rawls' *maximin* and of the utilitarian welfare function (i.e. summation of individual payoffs) (see [11, p. 851]).

Formally, consider two players i and j and let $x = \{x_i, x_j\}$ denote the vector of monetary payoffs. According to Charness & Rabin's model, the utility function of player i is given by:

$$U_i(x) = (1 - \lambda) \cdot x_i + \lambda \cdot [\delta \cdot min[x_i, x_j] + (1 - \delta) \cdot (x_i + x_j)]$$

where $\delta, \lambda \in [0, 1]$. Moreover, the two parameters can be interpreted as follows: δ measures the degree of concern for helping the worst-off person versus maximizing the total social surplus. Setting $\delta = 1$ corresponds to a pure "maximin" (or "Rawlsian" criterion), while setting $\delta = 0$ corresponds to total-surplus maximization. λ measures how much player i cares about pursuing the social welfare versus his own self-interest. Setting $\lambda = 1$ corresponds to purely "disinterested" preferences, in which players care no more (or less) about her own payoffs than others', while setting $\lambda = 0$ corresponds to pure self-interest.

As for the previous model, the parameters δ and λ can considerably change the structure of the ST game, which is why we propose a new game theoretical analysis involving such fair agents.

The first observation is that while fairness may slightly alter Bob's preferences, the (In, D_a, D_b) outcome always remains the best option: the only difference with the classical model is that he may come to prefer the (In, C_a, C_b) outcome to the (Out) solution when $\delta < 2/3$ and $\lambda > 1/3$.

Similarly, Alice's preferences also get affected by such notion of fairness. The main result is that a new forward induction solution may emerge through such a social preferences model:

- if $\lambda < 1/2$, then Alice may still play the forward induction solution as predicted by classical game theory (i.e. (In, C_a)), depending on the value of δ.

- if $1/2 \leq \lambda \leq 3/4$, then no prediction can be made without considering probabilistic beliefs: both Nash solutions in pure strategies from the subgame are always at least as good for Alice as playing (Out).
- if $\lambda > 3/4$ and $\delta > 2/3$, then Alice may play a forward induction solution (i.e. (In, D_a)) that mainly relies on her other regarding preferences: solution (In, D_a, D_b) indeed becomes preferred to playing Out, which is preferred to solution (In, C_a, C_b) (see Figure 4 for an example).

Moreover, one should note that, as for the original version of the game (see section 4), the (Out) option for Alice always dominates the Nash equilibrium in mixed strategies from the coordination subgame, no matter what the values of λ and δ are.

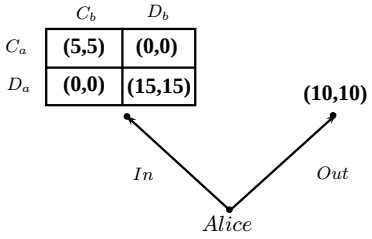


Figure 4. Transformed ST game for fair agents ($\lambda = \delta = 1$)

The above analysis suggests that the ST game may in fact contain two distinct focal points for the players, which can be identified by the two possible forward induction solutions. Therefore, one can state that the current ST game yields a unique social-welfare equilibrium[3] if and only if players have either some strong self-interested preferences ($\lambda << 1/5$) or some strong other-regarding preferences ($\lambda >> 3/4$ and $\delta >> 2/3$). In the latter case, one should note that the players' sensibility to the *maximin* principle needs to dominate that of the utilitarian welfare function.

The set of Nash Equilibria (NE), Subgame Perfect Equilibria (SPE), and Forward Induction solutions (FI), in the context of the ST game played by fair agents, is summarized through the following table:

NE	SPE	FI
$(In, D_a, D_b), (Out, C_a, C_b)$ (Out, D_a, C_b)	(In, D_a, D_b) (Out, C_a, C_b)	(In, D_a, D_b)

Table 2. Equilibrium solution concepts for fair agents ($\lambda >> 3/4$ and $\delta >> 2/3$)

6 Towards team reasoning

Another important concept that is of high relevance when studying social ties is about team reasoning. In fact, as already said in Section 2, players that are socially connected may be expected to identify themselves with the same group, which may consequently lead them to choose actions as a member of this group.

In order to illustrate this argument in the context of our ST game, let us define a payoff function U that satisfies, for example, Rawls' *maximin* criterion [29]. This criterion corresponds to giving infinitely greater weight to the benefits of the worse-off person. Applying this payoff function to the ST game leads to the transformed game depicted in Figure 4 from Section 5.2.

In fact, in this case, both players benefit if and only if they coordinate with each other in the subgame. However, their subsequent payoffs depends on which action they do coordinate on. The interesting property of this transformed subgame is that it introduces a dilemma that even economic theory cannot solve. However, while game theory is indeed unable to predict any particular outcome (i.e. both coordinated outcomes of the subgame are Nash solutions), it is shown in [2] that people would tend to coordinate on the action that leads to the most rewarding outcome for both (i.e. (D_a, D_b)). In order to interpret such intuitive behavior, some theorists have proposed to incorporate new modes of reasoning into game theory. For instance, starting from the work of Gilbert [20] and Reagan [30], some economists and logicians [26] have studied team reasoning as an alternative to the best-response reasoning assumed in classical game theory [33, 32, 1, 13]. Team-directed reasoning is the kind of reasoning that people use when they take themselves to be acting as members of a group or team [32]. That is, when an agent i engages in team reasoning, he identifies himself as a member of a group of agents S and conceives S as a unit of agency acting as a single entity in pursuit of some collective objective. A team reasoning player acts for the interest of his group by identifying a strategy profile that maximizes the collective payoff of the group, and then, if the maximizing strategy profile is unique, by choosing the action that forms a component of this strategy profile.

According to [22, 33], simple team reasoning (from Alice's viewpoint) in the current ST game can therefore be defined as follows:

Statement 6.1 *If Alice believes that:*

- *She is a member of a group* $\{Alice, Bob\}$.
- *It is common knowledge among Alice and Bob that both identify with* $\{Alice, Bob\}$.
- *It is common knowledge among Alice and Bob that both want the value of* U *to be maximized.*
- *It is common knowledge among Alice and Bob that* (In, D_a, D_b) *uniquely maximizes* U.

Then she should choose her strategy (In, D_a).

However, one should note that the above payoff function U is simply an example, and could then be interpreted otherwise. As an alternative, one may consider a function of social welfare that satisfies classical utilitarianism (i.e. by maximizing the total combined payoff of all players). In this case, as the transformed game would hold the same characteristics as the game depicted in Figure 4, Alice's behavior predicted by Statement 6.1 would remain unchanged.

7 Working hypotheses

As previously mentioned, the main goal of our ST game is to investigate whether social ties affect social preferences. According to the previous theoretical analyses, experimenting this game can therefore allow to verify the following hypotheses.

Hypothesis 7.1 *Social ties correlate with inequity aversion.*

Hypothesis 7.1 thus predicts that Alice will play (Out), no matter whether she is and/or expects Bob to be inequity averse.

Hypothesis 7.2 *Social ties correlate with fairness.*

Hypothesis 7.2 predicts that both Alice and Bob will coordinate on the (In, D_a, D_b) outcome. However, in this case, the following hypothesis also needs to be verified:

Hypothesis 7.3 *Social ties correlate with team reasoning*

[3] The social welfare equilibrium introduced by Charness & Rabin [11, p. 852] corresponds to a Nash equilibrium for some given values of δ and λ

Indeed, one should note that the ST game does not allow to distinguish Hypothesis 7.3 from Hypothesis 7.2 (in both cases, agents should play (In, D_a, D_b)). In order to differentiate these hypotheses, one may then consider a version of our game without the outside option (that is the possibility for Alice to play (Out) first): this simply corresponds to playing the coordination subgame alone. In this alternative situation, the game resembles the well known Hi-Lo matching game: Hypothesis 7.2 then predicts that players would miscoordinate (there will always be two different social welfare equilibria in this case), whereas Hypothesis 7.3 predicts that players would not change their behavior and still coordinate on the (D_a, D_b) outcome.

8 Conclusion

In this paper, we have proposed a game that appears to have very nice properties to investigate the behavioral effects of social ties. Indeed it creates a dilemma between maximizing self-interest and maximizing social welfare. It differs however from existing economic games from the literature that elicit similar properties, such as the trust game, the ultimatum game, and the dictator game. In the latter cases, both players only need to rely on their own type of preference as well as their belief about the other's, which may then be influenced by some psychological factors (e.g. disappointment, regret, guilt) [18]. On the other hand, in our ST game, knowing each other's type of preference is not sufficient to predict any action that maximizes utilities, it also needs to be common knowledge among them. In addition to allowing for some considerably more detailed epistemic analysis, such an additional constraint seems relevant as it appears to be a requirement for the existence of a social tie (according to Statement 2.1 from Section 2). Moreover, this game is also well suited to evaluate the very plausible theory of team reasoning in the context of social ties: the stronger the tie between individuals, the more they may act as members of the same group.

However, as this work is purely theoretical, it clearly suggests some further experimental analysis. The next stage of this study therefore consists of testing and evaluating the various hypotheses made in the previous sections. To do so, we intend to conduct experimental sessions where people will be asked to interact (1) with some perfect strangers, and (2) with some socially connected individuals (e.g. friends, class mates, team mates, etc…) in the context of our ST game in extensive form.

REFERENCES

[1] M. Bacharach, 'Interactive team reasoning: a contribution to the theory of cooperation', *Research in economics*, **23**, 117–147, (1999).

[2] M. Bacharach, *Beyond individual choice: teams and frames in game theory*, Princeton University Press, Oxford, 2006.

[3] D. Balkenborg and Sonderforschungsbereich 303-" Information und die Koordination wirtschaftlicher Aktivitäten.", *An experiment on forward versus backward induction*, Rheinische Friedrich-Wilhelms-Universität Bonn, 1994.

[4] J. Berg, J. Dickhaut, and K. McCabe, 'Trust, reciprocity, and social history', *Games and Economic Behavior*, **10**(1), 122–142, (1995).

[5] K. Binmore and L. Samuelson, 'Evolutionary drift and equilibrium selection', *The Review of Economic Studies*, **66**(2), 363, (1999).

[6] G. E. Bolton and A. Ockenfels, 'A theory of equity, reciprocity and competition', *America Economic Review*, **100**, 166–193, (2000).

[7] J. Brandts, A. Cabrales, and G. Charness, 'Forward induction and the excess capacity puzzle: An experimental investigation', (2003).

[8] J. Brandts and C.A. Holt, *Forward induction: Experimental evidence from two-stage games with complete information*, Departament d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona, 1989.

[9] J. Brandts and C.A. Holt, 'Limitations of dominance and forward induction: Experimental evidence', *Economics Letters*, **49**(4), 391–395, (1995).

[10] G.P. Cachon and C.F. Camerer, 'Loss-avoidance and forward induction in experimental coordination games', *The Quarterly Journal of Economics*, **111**(1), 165, (1996).

[11] G. B. Charness and M. Rabin, 'Understanding social preferences with simple tests', *Quarterly Journal of Economics*, **117**, 817–869, (2002).

[12] Y. Chen and S.X. Li, 'Group identity and social preferences', *The American Economic Review*, **99**(1), 431–457, (2009).

[13] A. M. Colman, B. N. Pulford, and J. Rose, 'Collective rationality in interactive decisions: evidence for team reasoning', *Acta Psychologica*, **128**, 387–397, (2008).

[14] R. Cooper, D.V. De Jong, R. Forsythe, and T.W. Ross, 'Forward induction in coordination games', *Economics Letters*, **40**(2), 167–172, (1992).

[15] R. Cooper, D.V. DeJong, R. Forsythe, and T.W. Ross, 'Forward induction in the battle-of-the-sexes games', *The American Economic Review*, 1303–1316, (1993).

[16] E. Fehr and K. M. Schmidt, 'A theory of fairness, competition, and cooperation', *Quarterly Journal of Economics*, **114**, 817–868, (1999).

[17] J.H. Frost, Z. Chance, M.I. Norton, and D. Ariely, 'People are experience goods: Improving online dating with virtual dates', *Journal of Interactive Marketing*, **22**(1), 51–61, (2008).

[18] J. Geanakoplos, D. Pearce, and E. Stacchetti, 'Psychological games and sequential rationality', *Games and Economic Behavior*, **1**(1), 60–79, (1989).

[19] G. Gigerenzer and R. Selten, 'The adaptive toolbox', *Bounded rationality: The adaptive toolbox*, 37–50, (2001).

[20] M. Gilbert, *On social facts*, Routledge, London, 1989.

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

[22] N. Gold and R. Sugden, 'Theories of team agency', (2007).

[23] W. Güth, R. Schmittberger, and B. Schwarze, 'An experimental analysis of ultimatum bargaining', *Journal of Economic Behavior & Organization*, **3**(4), 367–388, (1982).

[24] G.J. Hitsch, A. Hortaçsu, and D. Ariely, 'Matching and sorting in online dating', *The American Economic Review*, **100**(1), 130–163, (2010).

[25] S. Huck and W. Muller, 'Burning money and (pseudo) first-mover advantages: an experimental study on forward induction', *Games and Economic Behavior*, **51**(1), 109–127, (2005).

[26] Emiliano Lorini, 'From self-regarding to other-regarding agents in strategic games: a logical analysis', *Journal of Applied Non-Classical Logics*, **21**(3-4), 443–475, (2011).

[27] H. Margolis, *Selfishness, Altruism, and Rationality: A Theory of Social Choice*, University of Chicago Press, Chicago, 1982.

[28] M. Rabin, 'Incorporating fairness into game theory and economics', *American Econonic Review*, **83**(5), 1281–1302, (1993).

[29] J. Rawls, *A theory of justice*, Harvard University Press, Cambridge, 1971.

[30] D. Regan, *Utilitarianism and cooperation*, Clarendon Press, Oxford, 1980.

[31] Q. Shahriar, 'Forward induction works! an experimental study to test the robustness and the power', *Working Papers*, (2009).

[32] R. Sugden, 'Team preferences', *Economics and Philosophy*, **16**, 175–204, (2000).

[33] R. Sugden, 'The logic of team reasoning', *Philosophical Explorations*, **6**(3), 165–181, (2003).

[34] H. Tajfel, 'Experiments in intergroup discrimination', *Scientific American*, **223**(5), 96–102, (1970).

[35] H. Tajfel, M.G. Billig, R.P. Bundy, and C. Flament, 'Social categorization and intergroup behaviour', *European Journal of Social Psychology*, **1**(2), 149–178, (1971).

[36] B. Van Huyck John, C. Battalio Raymond, and O. Beil Richard, 'Asset markets as an equilibrium selection mechanism: Coordination failure, game form auctions, and tacit communication', *Games and Economic Behavior*, **5**(3), 485–504, (1993).

Conviviality by Design

Patrice Caire [1] and Antonis Bikakis [2] and Vasileios Efthymiou [3]

Abstract. With the pervasive development of socio-technical systems, such as Facebook, Twitter and digital cities, modelling and reasoning on social settings has acquired great significance. Hence, an independent soft objective of system design is to facilitate interactions. Conviviality has been introduced as a social science concept for multiagent systems to highlight soft qualitative requirements like user friendliness of systems. Roughly, more opportunity to work with other people increases the conviviality. In this paper, the question we address is how to design systems to increase conviviality by design. To evaluate conviviality, we model agent interactions using dependence networks, and define measures that quantify interdependence over time. To illustrate our approach we use a gaming example. Though, our methods can be applied similarly to any type of agent systems, which involve human or artificial agents cooperating to achieve their goals.

1 Introduction

As software systems gain in complexity and become more and more intertwined with the human social environment, models that can express the social characteristics of complex systems are increasingly needed [13, 8, 16]. For example, people may live far apart, speak different languages and have never physically met, but still, they expect to interact electronically with each other as they do physically. Hence, an implicit soft objective of system design is often to facilitate interactions. Conviviality emerges, but we want to design systems that foster conviviality among people or devices [18].

So far, most systems let users find their own ways to cooperate without providing any help or support. In such cases, users have to coordinate their actions and cooperate in a distributed way. Without any support from the system, they are not able to evaluate their cooperation and therefore the conviviality of the system; consequently they also cannot find ways to increase it. Conviviality is more than mere cooperation; it gives agents the freedom to chose with whom to cooperate.

Our proposed approach follows an alternative direction. It is based on the intuition that, to be convivial, the system itself should provide its users with potential ways to cooperate. For example, the system may suggest to the employees of a company, possible ways of interaction that will improve their cooperation. The system may monitor the evolution of these interactions, evaluate the agents' cooperation, and update the suggestions it makes to increase conviviality. Our research question is the following:

[1] University of Luxembourg, email: patrice.caire@uni.lu
[2] UCL, United Kingdom, email: a.bikakis@ucl.ac.uk
[3] University of Luxembourg, email: vasileios.efthymiou@uni.lu

Research Question: How to, by design, increase conviviality in multiagent systems?

This breaks down into the following sub-questions:

(a) How to evaluate conviviality?
(b) How to measure conviviality over time?
(c) What are the assumptions and requirements for such measures?
(d) How to use the measures in MAS?

In agent systems, conviviality measures quantify interdependence in social relations, representing the degree to which the system facilitates social interactions. Roughly, more interdependence increases conviviality among groups of agents or coalitions, whereas larger coalitions may decrease the efficiency or stability of these involved coalitions. We are, therefore, interested in two main issues. The first one is to design multiagent systems so that they foster conviviality, while the second one is to evaluate conviviality. For the first issue we adopt the paradigm of dependence networks, based on the intuition that conviviality may be represented by the interdependence among the agents of the system. For evaluating conviviality over time, we build on the *static* measures originally introduced in [4]. We extend these measures by proposing new ones, that we call *temporal* case.

In this paper, we build on the notion of social dependence introduced by Castelfranchi [7]. Castelfranchi brings concepts like groups and collectives from social theory to agent theory to enrich agent theory and develop experimental, conceptual and theoretical new instruments for social sciences.

Moreover, we take as a starting point the notion of dependence graphs and dependence networks initially elaborated by Conte and Sichman [20], and Conte et al. [21], and further developed by these authors [20].

We build on the *Temporal Dependence Networks*, introduced in [5] to compare time sequences of different dependence networks. This time however, we model the potential evolutions of sequences within the same dependence network. We introduce three principles to define three new measures, and therefore compare conviviality in Temporal Dependence Networks in a macro- and micro-organizational scale.

The remainder of the paper is structured as follows: First, we introduce our motivating example, highlighting the main challenges. Then, we identify requirements for convivial system design measures. We introduce our temporal dependence networks measures and principles. Finally, we present some of the most related works and summarize this paper.

2 Example Scenario

In order to demonstrate the requirements and challenges of conviviality among heterogeneous agents, we use an example scenario from the domain of social networks. This example allows us to compare different instances of a game and illustrate how the system may increase the conviviality by evaluating the games against a number of conviviality principles.

Consider a game in Facebook, in which different users form teams and cooperate in order to achieve a common goal. We assume the members of each team to be completely unknown to each other (they are not Facebook friends and they have no friends in common), and that the game allows only one-to-one interactions between team members. For the sake of simplicity, we also assume that each team must consist of the same number of players. The game consists in finding answers to questions involving information that is available in the public profiles of the team members. The game unfolds in three different phases, and for each phase there is one associated problem in the form of a question/answer to be solved.

For the first phase, the question ($Q1$) is: *"Which team member has the most in common with the others?"*. For example, in a five-member team *A: Alice, Bob, Carlo, Dimitra* and *Eve*, it could be that Eve has common interests with Alice in tennis, with Carlo in Spanish movies, and with Dimitra in ancient history. Alice and Dimitra have a common interest in climbing, and Bob and Carlo are both interested in football. For team A, the correct answer would be 'Eve".

The second phase question ($Q2$) is: *"Which country corresponds to both the picture uploaded by answer of Q1* (Eve) *and one (and one only) of the team members?"*. For our team A, the correct answer would be "Greece" based on the fact that Eve has uploaded a photo, which was taken in Athens, and Dimitra is the only team member that comes from Greece.

The last question ($Q3$) is: *"What is the place among the answers provided to Q2 that most team members prefer?* (Greece). The answer would be "Santorini", which is "liked" by Alice, Dimitra and Eve, while other places in Greece, such as Athens or Crete, are "liked" only by two of the team members.

The team that manages to solve the riddles faster than the other teams is the winner. Building on instances of the game, we analyze how the system may increase the conviviality of the game by evaluating it against proposed principles.

Winning such a game requires finding the proper ways to cooperate, and assessing the team's performance by evaluating conviviality. In brief, the challenges of this game are:

1. **Cooperation.** If one of the team members does not cooperate, this would probably mean that the team may not be able to answer a question, and consequently win the game. The challenge, here, is to enable and foster cooperation between the team players.
2. **Evaluation of conviviality.** This process will help the team assess its performance in each round of the game, and find ways to improve it. For example, if team A could not provide an answer to $Q1$, because there were not enough interactions between the team members, the team should be able to realise the reasons for their poor performance and find ways to improve it for the next rounds. The challenge, in this case, is to develop principled methods for measuring the conviviality among the team members.

3 Hypotheses and requirements

To represent agents' interdependencies we use dependence networks [9, 19, 2], differentiating static and temporal cases.

3.1 Static case

In this case, all interdependencies are modelled in a single "global" dependence network, as in [9, 19, 2]. We consider that the agents' goals and interdependencies have been identified using a goal-oriented method like Tropos [3], for instance. Abstracting from method-specific concepts (e.g. tasks and resources in Tropos), we define a dependence network as in [4]:

Definition 3.1 (Dependence network) *A dependence network (DN) is a tuple $\langle A, G, dep, \geq \rangle$ where: A is a set of agents, G is a set of goals, $dep : A \times A \rightarrow 2^G$ is a function that relates with each pair of agents, the sets of goals on which the first agent depends on the second, and $\geq$: $A \rightarrow 2^G \times 2^G$ is for each agent a total pre-order on sets of goals occurring in his dependencies: $G_1 >_{(a)} G_2$.*

To illustrate our definition, we consider that during the first phase of the game, only A and B interact to answer Q1; during phase 2, B and C interact as well as D and C; and during phase 3, B and E interact as well as D and E, and A and E. Figure 1 depicts a dependence network that captures this situation. The nodes A, B, C, D and E represent agents Alice, Bob, Carlo, Dimitra and Eve. The arrows indicate the goal dependencies (i.e. ask a question or reply to it). A number of coalitions are formed among the five agents, such as (A, E), (A, B, E) and (A, B, C, D, E).

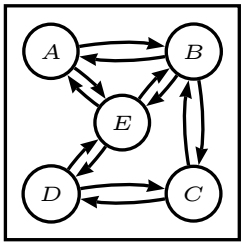


Figure 1. Example of a dependence network.

Based on [4], we make the following hypotheses:

- H1 the cycles identified in a dependence network are considered as coalitions. These coalitions are used to evaluate conviviality in the network. Cycles are the smallest graph topology expressing interdependence, thereby conviviality, and are therefore considered atomic relations of interdependence. When referring to *cycles*, we are implicitly signifying *simple cycles*, i.e., where all nodes are distinct [10]; we also discard self-loops. When referring to conviviality, we always refer to potential interaction not actual interaction.
- H2 conviviality in a dependence network is evaluated in a bounded domain, i.e., over a $[0, 1]$ interval. This allows the comparison of different systems in terms of conviviality.
- H3 larger coalitions have more conviviality.
- H4 the more coalitions in the dependence network, the higher the conviviality measure (ceteris paribus).

Our top goal is to maximize conviviality in the multiagent system. Some coalitions provide more opportunities for their participants to cooperate than others, being thereby more convivial. Our two sub-goals (or requirements) are thus:

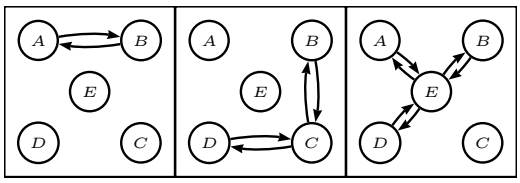
R1 maximize the size of the agent's coalitions, i.e. to maximize the number of agents involved in the coalitions,
R2 maximize the number of these coalitions.

3.2 Temporal Case

For more fine-grained exploration, the network can be divided up into sequences, and analysis performed on each sequence. This allows for local analysis of the network and is less computationally intensive. Definition 3.2 formalizes how dependence networks can be extended to capture the temporal evolution of dependencies between agents, inspired from [5].

Definition 3.2 (Temporal dependence network) *A temporal dependence network (TDN) is a tuple ⟨A,G,T, dep⟩ where: A is a set of agents, G is a set of goals, T is a set of natural numbers denoting the time units or sequence number, $dep : T \times A \times A \rightarrow 2^G$ is a function that relates with each triple of a sequence number, and two agents, the set of goals on which the first agent depends on the second.*

Returning to our example, the static view illustrated Figure 1 is now captured as a sequence in Figure 2. If we call the temporal dependence network TDN_k, TDN_k^j denotes the individual dependence network that corresponds to the j^{th} step. Note that $|A|$, the number of agents (5 in this case), remains constant over TDN_k. $|TDN_k|$ refers to the length of the temporal dependence network (3 in this case).



(a) TDN_k^1 (b) TDN_k^2 (c) TDN_k^3

Figure 2. Example of a temporal dependence network.

Building on the static case, our assumptions are:

H5 the more regularly the number of coalitions increases, the higher the conviviality measure (ceteris paribus); for example, in human society, allowing people to get to know each other progressively enables trust to build up. In cases, where agents need to quickly form a grand coalition without build up, and dissolve, the assumptions may differ.
H6 the more different agents take part in coalitions, the higher the conviviality (ceteris paribus); for example, by allowing all agents to participate in interactions.

Our two additional requirements are thus:

R3 maximize the regular increment of the number of coalitions,
R4 maximize the involvement of each individual agent in the coalitions.

4 Conviviality measures

In multiagent systems, conviviality has been evaluated by measuring the interdependencies among the agents [4]. In this section, we use the static conviviality measures presented in [4], that we call *static* case. We extend these measures by proposing new ones, that we call *temporal* case. The main challenge in defining conviviality measures over time is to make assumptions about the sequences. For example, when modelling the agents' interdependencies as a sequence of dependence networks, we could leave out one dependence network from a sequence, or introduce multiple copies of the same dependence network. How this affects the conviviality and its evaluation depends on the underlying assumptions.

4.1 Static Case

The basic idea for the conviviality measures introduced in [4], is the following. Since the atomic structure reflecting conviviality is a pair of reciprocating agents, the conviviality measures should also be based on the pairing relations in the dependence networks. Hence, for each pair of agents, the number of cycles that contains this pair is counted. Furthermore, the measures introduced in [4] were normalized to be in $[0, 1]$ in order to allow the sensible comparison of any two dependence networks in terms of conviviality. Equation 1 is the general formula to express the pairwise conviviality measure $conv(DN)$ of a dependence network.

$$conv(DN) \quad = \quad \frac{\sum coal(a,b)}{\Omega}, \tag{1}$$

where $\text{coal}(a, b)$ for any distinct $a, b \in A$ is the number of cycles that contain both a and b in DN and Ω is the maximum the sum in the numerator can get, over a dependence network of the same set of goals and the same number of agents but with all possible dependencies.

To compare the conviviality of each of the three steps in TDN_k of Figure 2, using the measure of Equation 1, we would just have to count the pairs of agents that belong to cycles, since the denominator Ω is the same for all three steps. In TDN_k^1 there are two pairs participating in a cycle: $(A, B), (B, A)$, in TDN_k^2, four pairs of agents: $(B, C), (C, B), (C, D), (D, C)$ and in TDN_k^3 six pairs: $(A, E), (E, A), (B, E), (E, B), (D, E), (E, D)$. This makes the third step more convivial than the first two.

4.2 Temporal Case

Conviviality in Temporal Dependence Network can be measured on at least two separate scales: the micro organizational and the macro-organizational scales. Measurements at the macro-organizational scale focus on the evaluation and comparison of the conviviality measures of each step in the sequence of dependence networks, whereas micro-organizational measurement reflects topological aspects within each dependence network. We consider three measurement principles:

Principle 1 (Dominance) *A temporal dependence network has more conviviality than another one if, ceteris paribus, each individual dependence network of the former has more conviviality than the corresponding (same sequence number) individual dependence network of the latter. This is a combination of R1 and R2 from the single transition case.*

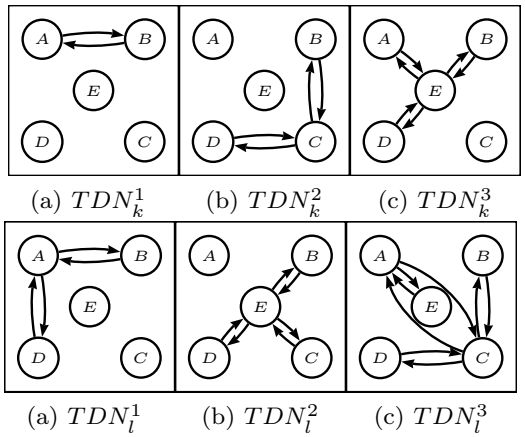
Principle 2 (Volatility) *A temporal dependence network has more conviviality than another one if, ceteris paribus, the conviviality measures of all individual dependence networks in the former shows less volatility than in the latter.*

Principle 3 ((Micro-organizational) Entropy) *A temporal dependence network has higher conviviality than another one if, ceteris paribus, the dependence topology in the former shows more variations than in the latter, i.e., if the agents have the opportunity to interact in a greater variety of coalitions.*

For instance, when we state our Principle 1, *Dominance*, we compare conviviality measures of each step in the sequence of dependence networks, thus a measure at the macro-organizational is done. The same holds when we say that the conviviality measures should be equally distributed (Principle 2, *Volatility*). In contrast, to be able to compare the entropy within two sequences of temporal dependence networks, and evaluate the R.4, i.e., maximize the involvement of each individual agent in the coalitions, we need to study the temporal dependence network at a micro-organizational scale.

4.2.1 Macro-organizational scale

To illustrate our *Dominance* Principle, we return to our running example. Consider two instances of the game: l and k. The same five players, Alice, Bob, Carlo, Dimitra and Eve, are trying to improve their conviviality. Indeed, in game l they considered that they did poorly. They play a second game k and compare their performance with the first one. Figure 3 illustrates the *Dominance* Principle with these two games.



(a) TDN_k^1 (b) TDN_k^2 (c) TDN_k^3

(a) TDN_l^1 (b) TDN_l^2 (c) TDN_l^3

Figure 3. Illustration of Dominance.

The first game l, represented by the temporal dependence network TDN_l has more conviviality than the second, represented by TDN_k. In each corresponding phase of the game, there are more interactions among the agents in game l than in game k. For example, in phase 1, three agents from game l interact, namely A, D and B, to form two coalitions, whereas in the same phase, only two agents from game k interact, namely A and B, to form a single coalition.

We now introduce our fine-grained conviviality measures for temporal dependence networks. Let TDN_1 and TDN_2 be two temporal dependence networks.

Let $|TDN_1|$ and $|TDN_2|$ be the length of these temporal dependence networks, i.e., the number of steps in the sequences. Let $|A_1|$ and $|A_2|$ be the number of agents in TDN_1 and TDN_2 respectively. We recall that $|A_1|$ and $|A_2|$ are constant over the individual dependence networks Let TDN_i^j denote the j-th individual dependence network of the temporal dependence networks TDN_i.

Definition 4.1 (Dominance, formally) *Let* $|TDN_1| = |TDN_2|$. *If* $\forall TDN_1^j$ $conv(TDN_1^j) \geq conv(TDN_2^j)$, *then* $conv(TDN_1) \geq conv(TDN_2)$.

For each instance of TDN_l in Figure 3, the corresponding instance of TDN_k, containing the same agents and goals, has less cycles. This makes TDN_l overally more convivial.

Similarly as in the static case represented Figure 1, we can assume, for our example, that each cycle consists of the same two goals reciprocation in any given individual dependence network. For instance, illustrated Figure 3, in TDN_k^2, C depends on B and reciprocally, to ask and answer question, similarly C depends on D and reciprocally. This reflects the fact that the game is turn based, and all players have similar goals at a given phase of the game (i.e., in a given individual dependence network step). Then, there are a total of 2 goals in each individual dependence network of our examples (Figure 3 to Figure 5). The following are then constant over all the computation section for each individual dependence network:

- $Agents = \{A, B, C, D, E\}$,
- $Goals = \{$"ask a question", "reply to a question"$\}$,
- $\Omega = 6320$.

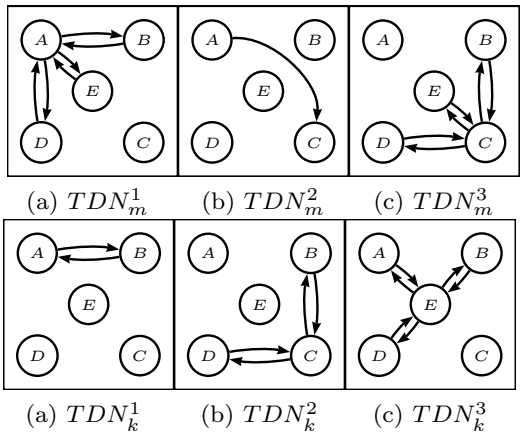
The conviviality computation of each individual dependence network step displayed on Figure 3 is presented in Table 1. For instance, the conviviality of TDN_k is explained in Paragraph 4.1. We see that the computed conviviality for each individual dependence network is higher in TDN_l than in TDN_k. In each phase of the game, the players have more interactions.As a conclusion and per *Dominance* Principle, TDN_l has more conviviality than TDN_k.

Table 1. Computations for TDN_k and TDN_l.

Phase 1	Phase 2	Phase 3
$conv(TDN_k^1) = \frac{2}{\Omega}$	$conv(TDN_k^2) = \frac{4}{\Omega}$	$conv(TDN_k^3) = \frac{6}{\Omega}$
$conv(TDN_l^1) = \frac{4}{\Omega}$	$conv(TDN_l^2) = \frac{6}{\Omega}$	$conv(TDN_l^3) = \frac{8}{\Omega}$

We illustrate our second Principle *Volatility*, corresponding to our Requirement R3, by comparing a previous instance of the game, namely k with a new one m, in which agents have had the same number of interactions to answer Q1 in phase 1 and Q3 in phase 3, but no reciprocal interaction to address Q2 in phase two. Figure 4 illustrates this case. The temporal dependence network TDN_k has more conviviality than TDN_m. In game k, players change their interactions more gradually over the three phases, whereas changes in game m are more erratic, going from many interactions in phase 1 to no interaction in phase 2, to many interactions again in phase 3.

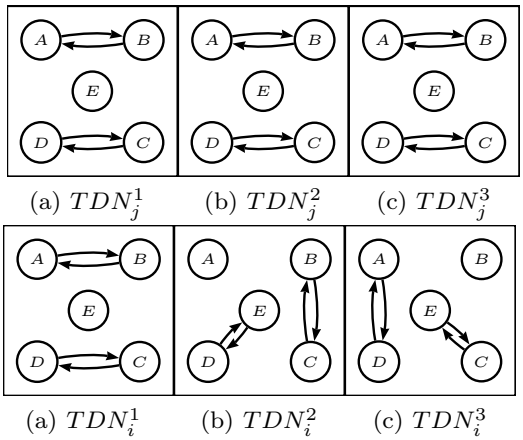
We use the notion of standard deviation σ, which reflects the volatility in a set of measures. A low standard deviation indicates that data points tend to be very close to the mean, whereas high standard deviation indicates that the data is



(a) TDN_m^1 (b) TDN_m^2 (c) TDN_m^3

(a) TDN_k^1 (b) TDN_k^2 (c) TDN_k^3

Figure 4. Illustration of Volatility.



(a) TDN_j^1 (b) TDN_j^2 (c) TDN_j^3

(a) TDN_i^1 (b) TDN_i^2 (c) TDN_i^3

Figure 5. Illustration of Entropy.

spread out over a large range of values. We note $\sigma(TDN_i)$ the standard deviation over the individual dependence networks belonging to the temporal dependence network TDN_i. We also need to fix the conviviality mean of TDN_1 and TDN_2, respectively noted $\mu(TDN_1)$ and $\mu(TDN_2)$.

Definition 4.2 (Volatility, formally) *Let* $|TDN_1| = |TDN_2|$, *and* $\mu(TDN_1) = \mu(TDN_2)$. *If* $\sigma(TDN_1) < \sigma(TDN_2)$, *then* $conv(TDN_1) > conv(TDN_2)$.

Even if the two temporal dependence networks of Figure 4 have the same mean value for conviviality, $\frac{4}{\Omega}$, the standard variation of TDN_k is less than the standard variation of TDN_m. This means that the conviviality of TDN_k changes more gradually and therefore TDN_k is more convivial. The intuition for this principle is that volatility and dependency are two conflicting notions.

To evaluate the conviviality of the temporal dependence networks depicted Fig. 4, we first compute conviviality for each individual dependence network step, presented Table 2.

Table 2. Computations for TDN_m and TDN_k, Fig. 4.

Phase 1	Phase 2	Phase 3
$TDN_m^1 = \frac{6}{\Omega}$	$TDN_m^2 = 0$	$TDN_m^3 = \frac{6}{\Omega}$
$TDN_k^1 = \frac{2}{\Omega}$	$TDN_k^2 = \frac{4}{\Omega}$	$TDN_k^3 = \frac{6}{\Omega}$

Table 3 presents the means and the standard distribution, showing that TDN_k is more convivial than TDN_m, as $\sigma(TDN_m) > \sigma(TDN_k)$.

Table 3. Means and standard distribution.

	Game m	Game k
Means	$\mu(TDN_m) = \frac{4}{\Omega}$	$\mu(TDN_k) = \frac{4}{\Omega}$
St. dist.	$\sigma(TDN_m) = \sqrt{\frac{8}{\Omega^2}}$	$\sigma(TDN_k) = \sqrt{\frac{8}{3\times\Omega^2}}$

4.2.2 Micro-Organizational Scale

Figure 5 illustrates *Entropy*: TDN_i is more convivial than TDN_j. In game i, players change partners more often, allowing all players to interact, whereas in game j the same players interact with each other and one player is never involved.

Let δ_T be the number of different coalitions over all steps in the sequences of the temporal dependence network T.

Definition 4.3 (Entropy, formally) *Let* $|TDN_1| = |TDN_2|$, *and* $\mu(TDN_1) = \mu(TDN_2)$, *and* $\sigma(TDN_1) = \sigma(TDN_2)$. *If* $\delta_1 > \delta_2$, *then* $coal(TDN_1) > coal(TDN_2)$.

In Figure 5, none of the two temporal dependence networks TDN_j and TDN_i is dominant or less volatile. However, in TDN_j the same coalitions exist throughout the game, whereas in TDN_i, different coalitions are formed and consequently more players have the ability to participate, contribute and benefit. Therefore, TDN_i is more convivial.

Table 4. Entropy, Fig. 5.

$\mu(TDN_j) = \frac{4}{\Omega}$	$\sigma(TDN_j) = 0$	$\delta_{TDN_j} = 2$
$\mu(TDN_i) = \frac{4}{\Omega}$	$\sigma(TDN_i) = 0$	$\delta_{TDN_i} = 6$

Remark: this principle may lead to unexpected results since only the number of coalitions is taken into account (and not their length). If we limit ourself to coalitions of length 2, the above is sufficient. A further study is needed to understand the impact of this principle on coalitions with random lengths.

4.2.3 Discussion

In this section we define conviviality measures that satisfy the four requirements we distinguish and the three principles for our conviviality measures, and illustrate them with our running example. Our measures build up to allow the agents to compare their performances and increase their conviviality. Our first measures allow agents to compare their conviviality at each step of the game. However, these measures do not reflect the distribution of conviviality over the whole sequence, which is what our second measures provide. On the other hand, these second measures do not provide any insight on which agents cooperate to ensure individual agents' participation, which is addressed by our third measure.

5 Related research

In this paper, we use the notion of social dependence introduced by Castelfranchi [7]. Moreover, we build on the notion of dependence graphs and dependence networks, elaborated by Conte and Sichman [20], and Conte et al. [21], in order to model and measure conviviality.

By contrast, we use a more abstract representation of dependence networks, i.e., abstracting notions such as tasks, actions or plans. In this sense our approach also builds to Sauro's abstractions in [15], Boella et al. [2].Dependence based coalition formation is analyzed by Sichman [19], while other approaches are developed in [17, 11, 1].

Differently from Grossi and Turrini [12], our approach does not bring together coalitional theory and dependence theory in the study of social cooperation within multiagent systems. Moreover, our approach differs as it does not hinge on agreements. Finally, similarly to works such as in Johnson and Bradshaw et al. "coactive" design [14], we emphasize agents' interdependence as a critical feature of multiagent systems. Addtionally, the authors focus on the design of systems involving joint interaction among human-agent systems .

6 Summary

In agents systems, conviviality measures quantify interdependence in social dependence relations, representing the degree to which the system facilitates social interactions. In this paper, we distinguish static from temporal measures. In the static case, roughly, more interdependence increases conviviality among groups of agents, i.e., coalitions, whereas larger coalitions may decrease the efficiency or stability of these involved coalitions. In the temporal case, we consider sequences of dependence networks over time.

We distinguish four requirements to maximize conviviality in a multiagent system: 1) maximize the size of the agent's coalitions; 2) maximize the number of these coalitions; 3) maximize the regular increment of the number of coalitions; and 4) maximize the involvement of each individual agent in the coalitions. Furthermore, we distinguish three principles to guide our definition of conviviality measures: **dominance**, **volatility**, and **entropy**. Finally, we define conviviality measures that can be used to test our requirements following our three principles, and illustrate them with a gaming example.

A topic of further work is to define measures of temporal dependence networks for other interpretation of the temporal sequence, and to define conviviality measures for dynamic normative dependence networks. The difference between the two, is that in the latter, a normative system mechanism is used to change conviviality by changing social dependencies, for example by creating new obligations, hiding power relations and social structures. This has been used to define conviviality masks [6], and thus the measures of dynamic dependence networks will lead to measures of conviviality masks. However, we expect that the proposed measures do not apply in a straightforward way, but that new measures will be needed to capture further views of conviviality.

REFERENCES

[1] G. Boella, L. Sauro, and L. van der Torre. Algorithms for finding coalitions exploiting a new reciprocity condition. *Logic Journal of the IGPL*, 17(3):273–297, 2009.

[2] G. Boella, L. Sauro, and L. W. N. van der Torre. Power and dependence relations in groups of agents. In *IAT*, pages 246–252. IEEE Computer Society, 2004.

[3] P. Bresciani, A. Perini, P. Giorgini, F. Giunchiglia, and J. Mylopoulos. Tropos: An agent-oriented software development methodology. *Autonomous Agents and Multi-Agent Systems*, 8(3):203–236, 2004.

[4] P. Caire, B. Alcade, L. van der Torre, and C. Sombattheera. Conviviality measures. In *10th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011), Taipei, Taiwan, May 2-6, 2011*, 2011.

[5] P. Caire and L. van der Torre. Temporal dependence networks for the design of convivial multiagent systems. In *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, May 10-15, 2009, Volume 2*, pages 1317–1318, 2009.

[6] P. Caire, S. Villata, G. Boella, and L. van der Torre. Conviviality masks in multiagent systems. In *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, May 12-16, 2008, Volume 3*, pages 1265–1268, 2008.

[7] C. Castelfranchi. The micro-macro constitution of power. *Protosociology*, 18:208–269, 2003.

[8] R. Conte, M. Paolucci, and J. Sabater Mir. Reputation for innovating social networks. *Advances in Complex Systems*, 11(2):303–320, 2008.

[9] R. Conte and J. Sichman. Dependence graphs: Dependence within and between groups. *Computational and Mathematical Organization Theory*, 8(2):87–112, July 2002.

[10] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2nd edition, 2001.

[11] A. Gerber and M. Klusch. Forming dynamic coalitions of rational agents by use of the dcf-s scheme. In *AAMAS*, pages 994–995, 2003.

[12] D. Grossi and P. Turrini. Dependence theory via game theory. In W. van der Hoek, G. A. Kaminka, Y. Lespérance, M. Luck, and S. Sen, editors, *AAMAS*, pages 1147–1154. IFAAMAS, 2010.

[13] A. Haddadi. *Communication and Cooperation in Agent Systems, A Pragmatic Theory*, volume 1056 of *Lecture Notes in Computer Science*. Springer, 1995.

[14] M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. Sierhuis, and B. van Riemsdijk. Toward coactivity. In P. J. Hinds, H. Ishiguro, T. Kanda, and P. H. K. Jr., editors, *HRI*, pages 101–102. ACM, 2010.

[15] L. Sauro. *Formalizing Admissibility Criteria in Coalition Formation among Goal Directed Agents*. PhD thesis, University of Turin, Italy, 2006.

[16] M. Seredynski, P. Bouvry, and M. A. Klopotek. Preventing selfish behavior in ad hoc networks. In *IEEE Congress on Evolutionary Computation*, pages 3554–3560. IEEE, 2007.

[17] O. Shehory and S. Kraus. Methods for task allocation via agent coalition formation. *Artif. Intell.*, 101(1-2):165–200, 1998.

[18] Y. Shoham and M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies (preliminary report). In *AAAI*, pages 276–281, 1992.

[19] J. S. Sichman. Depint: Dependence-based coalition formation in an open multi-agent scenario. *J. Artificial Societies and Social Simulation*, 1(2), 1998.

[20] J. S. Sichman and R. Conte. Multi-agent dependence by dependence graphs. In *Procs. of The First Int. Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2002*, pages 483–490. ACM, 2002.

[21] J. S. Sichman, R. Conte, C. Castelfranchi, and Y. Demazeau. A social reasoning mechanism based on dependence networks. In *ECAI*, pages 188–192, 1994.

Conformist imitation, normative agents and Brandom's commitment model

Rodger Kibble[1]

Abstract. This paper focuses on the role of imitation in social learning and everyday interaction, and proposes the outline of a framework based on a modified version of Robert Brandom's model of doxastic (propositional) and practical commitments. We question Brandom's assumption that there is a fundamental asymmetry between these two types of commitment and argue that conformist imitation can be incorporated into his model if we allow that practical as well as propositional commitments may be accorded default entitlement and that (provisional) entitlement may be inherited from other agents. Thus alongside Brandom's notion of inheritance of entitlement to propositional commitments via testimony, we propose inheritance by example in the practical case. This line of argument is contrasted with recent computational models based on data mining and machine learning. Finally, we briefly discuss how these findings may be incorporated in a framework for normative agents.

1 INTRODUCTION

A recent survey of the state of the art in normative multi-agent systems [12] proposes a model of the "norm life cycle" incorporating the processes of creation, transmission, recognition, enforcement, acceptance, modification, internalisation, emergence, forgetting and evolution. This paper will focus on one particular aspect of social learning and interaction, namely conformist imitation, and will suggest ways it can be incorporated into this model.

Imitation has been called "the main process of social learning" [16] and there is evidence that the propensity to imitate is one of the key factors distinguishing humans from other higher primates, along with productive use of language and large-scale cooperation outside kin groups [10]. The field of agent-based social simulation has taken on board the notion of social learning from social psychology: there has been much discussion of agents' propensity to imitate others in learning and interaction [16, 7]. [10] marshals evidence that a disposition to imitate may in fact be "hard wired" in humans:

> In the same way that individuals develop certain responsive dispositions, which lead them to develop appropriate beliefs in the case of observations, or desires in the case of somatic stimulus, people also acquire rules to govern their conduct by imitating observed regularities of behaviour in their immediate social environment.

Furthermore, the choice of which behaviour to imitate is subject to a "conformist bias": if there are competing regularities in a population, individuals will tend to select the one which is most common.

[1] Department of Computing, Goldsmiths University of London. Email: r.kibble@gold.ac.uk

[12] distinguishes between Type I norms, which are decreed by an authority, and Type II which emerge from interactions between agents. I would like to distinguish further between two classes of Type II norms: what we may call *behaviourist* norms, essentially regularities in behaviour governed by positive or negative reinforcement, and *intersubjective* norms which are characterised by mutual accountability between agents. Thus for example if someone takes it on themselves to sanction an "incorrect" action, their entitlement to carry out sanctions is itself at issue. The aim of this paper will be to show how conformist imitation can be accounted for within an intersubjective normative framework.

My main thesis will be that imitation is a manifestation of inherited entitlement to practical commitments as defined in Robert Brandom's account of normativity [2, 3]. The account will be based on Brandom's commitment model but will argue for some significant modifications to his framework. The remainder of this paper will be structured as follows. Section 2 will draw a distinction between instrumental accounts of normativity and approaches based on essentially communicative models of rationality, involving notions such as accountability and justification. This distinction will be motivated via critical discussion of some recent proposals in the field of agent-based modelling. Section 3 will outline some essential characteristics of Brandom's commitment model, while section 4 will propose detailed arguments for default entitlement and inheritance of entitlement to practical commitments. Section 5 will sketch possible applications to normative MAS architectures and section 6 presents some concluding remarks.

2 NORMS VERSUS REGULARITIES

Is there a clear distinction between norms and regularities? By "norm", I mean here a type of behaviour towards which it is appropriate to take a normative stance: that is, the behaviour is generally approved, and it is considered appropriate both to sanction those who breach the norm and those who fail to sanction non-compliance. 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. While it is clear that imitative behaviour can lead to regularities, it is perhaps less clear that it can establish norms. This construal of "norms" turns out to be quite similar to 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". And indeed it seems quite plausible that

a given practice can be considered to be "correct" within a community without any members of the community being able to quantify how frequently this practice is observed.

2.1 Where do norms come from?

The survey referred to above [12] cites two recent studies [20, 19] as exemplars of agent-based simulations which aim to model the emergence or acquisition of what I have called "behaviourist" norms. [20] treats norm emergence as a problem of resolving social dilemmas where there are multiple game-theoretic equilibria. The particular scenario investigated is the emergence of "rules of the road", i.e. whether to drive on the left or the right. The set-up is that when two drivers meet on the same side of the road, they have the options of both driving on (and colliding), both stopping, or one yielding to the other. Simulations involving various learning algorithms show that a population can converge on a convention to drive on one side or the other through multiple repeated interactions. The authors quote Axelrod on the self-enforcing nature of norms: "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". However, the "rules of the road" scenario doesn't fit this 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 the driver collides with an oncoming vehicle or stops because his way is blocked; and these consequences are equally costly for the conformist and the deviant. And it really seems to make little sense to talk of sanctions during the period of emergence of the putative norm, since one can only speak of conformists and deviants (and thus of appropriate use of sanctions) once the norm is in place.

[19] 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. The procedure the diner follows is:

1. Observe a series of episodes, some of which include sanctioning actions and some do not
2. Apply data mining techniques to discover if the sanctioning action is reliably associated with the presence of absence of any identifiable sequence of events.
3. Compile a set of candidate norms, namely regularities in behaviour which appear to be associated with sanctions.
4. Ask another agent in the vicinity whether a candidate norm is in fact a norm of the society. If the agent responds positively, the agent infers that the identified action is governed by an obligation norm. This is the "norm identification" stage.

In this scenario, the sanctioned action might be failure to leave a tip at the end of the meal, with the sanctioning action being some expression of disapproval or anger by the waiter. Thus, the diner's goal is to imitate the behaviour of other agents who are more successful in that they avoid being punished. The authors present simulation results showing the effect on the uptake of norms brought about by varying parameters such as the length of the event history that the diner takes into account, or a probability threshold for identifying candidate norms. Under certain assumptions the system does indeed succeed in learning that tipping is expected. Now, this process does fall a little short of "norm recognition": at best the system recognises candidate norms, which then have to be "verified" by asking a local (another agent in the vicinity). It could be argued that what the diner has identified is not a full-fledged norm but rather a regularity: when customers fail to leave tips, waiters are disposed to sanction them. 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. Short of physical violence, it is not always obvious to strangers whether particular actions count as friendly or hostile. 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. And once the diner conjectures that non-tipping may be meant as a sanction, he will have to observe several episodes in order to establish what kind of behaviour on the waiter's part is being punished. This observation might have to take account not only of sequences of events but e.g. the time that elapses between events. If a customer has failed to leave a tip because he has good reason to be unsatisfied with the service, then it may not be appropriate for the waiter to sanction him.

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. 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.

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 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. The fact that an unexplained and problematic notion of "sanctioning events" is used to "explain" norm identification may appear to be a fatal flaw in the proposal, or one could see it as pointing towards a deeper issue: normative frameworks may turn out to be unavoidably holistic and non-well-founded, only explicable in terms of other norms.

The arguments presented in this section are not particularly novel but draw on philosophical critiques of "regulist" and "regularist" approaches to normativity [2, 18]. Regulism corresponds to Type I above and construes norms as explicit rules or precepts laid down and enforced by some authority. Regularism corresponds to what I have called the "behaviourist" variant of Type II, according to which norms are quantifiable regularities in the behaviour of members of a community which are reinforced by positive or negative sanctions. Brandom [2] argues that both notions are incoherent and prone to infinite regress. 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: just as for instance a system of logical axioms is inert without some system of inference rules defining how the axioms are to be used in constructing proofs. This, it is ar-

gued, gives rise to a regress which must bottom out in rules that are implicit in practice. 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". Regularism also runs into a regress problem since as I show above, sanctioning is also a norm-governed activity which may be done correctly or incorrectly. Brandom and Rouse further accuse regularists of what Brandom calls "gerrymandering": the claim is that there is no uniquely identifiable sequence of actions that make up a norm-conformative performance. For example, it might happen that all the non-tippers in a restaurant scenario were wearing white socks, and that this was the cause of the waiter's ire. To be honest, this argument has the air of armchair theorising: 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. However, the criticism does seem valid for the particular model presented by [19]. The repertoire of actions is limited to a rather basic set comprising {arrive, order, eat, pay, tip, depart} for customers and {wait, sanction} for waiters: thus it is assumed that agents only perceive actions which are directly relevant to the problem under analysis. Indeed, the diner is assumed to be already equipped with the notion of "tipping", which puts in question whether this model could be extended to cover the acquisition of norms which are completely outside the agent's prior experience.

3 BRANDOM'S COMMITMENT MODEL

I have argued that norm-conformant behaviour such as conformist imitation is best modelled within a framework of mutual accountability, such that agents are in principle capable of questioning and justifying each others' behaviour. The remainder of this paper aims to provide an outline account within Robert Brandom's normative pragmatics, which uses parallel notions of social commitments and entitlements to model on the one hand actions and intentions, and on the other, assertions and beliefs. [13] rehearsed some classic issues with the BDI framework for multi-agent communication, derived in part from Austin and Searle's Speech Act theories, and proposed that Brandom's normative framework might form the basis of a more manageable approach. Brandom's approach is concerned with "deontic" attitudes of hearers, and of speakers as self-monitors, rather than intentional attitudes of speakers as in classic Speech Act theory. In place of beliefs and desires, Brandom discusses "doxastic" (propositional) and practical commitments, which interacting agents may acknowledge or ascribe to one another.

The normative dimensions of language use according to Brandom comprise *responsibility* - if I make a claim, I am obliged to back it up with appropriate evidence, argumentation and so on - and *authority* - by making a claim to which I am assumed to be entitled, I license others to make the same claim. Concepts are essentially rules or norms which govern the inferences we may or must make. The essential idea is that making an assertion is taking on a commitment to defend that assertion if challenged. There are obvious shared concerns with the notions of commitment developed by [9, 23] and introduced into MAS by [21]. Brandom's elaborations include the notion of *entitlement* to commitments by virtue of evidence, argumentation etc; the interpersonal inheritance of commitments and entitlements, and the treatment of consequential commitments and incompatibility

The mechanism for keeping track of agents' commitments and entitlements consists of deontic scoreboards maintained by each interlocutor, which record the set of commitments and entitlements which agents claim, acknowledge and attribute to one another (claims and acknowledgements are forms of self-attribution). Scoreboards are perspectival and may include both explicitly claimed commitments and consequential commitments derived by inference. Thus an agent may be assessed by others as being committed to propositions which are entailed by his overt commitments, whether or not he acknowledges such commitments. Agents may be in a position of claiming incompatible commitments but may not be assessed as entitled to more than one of them (if any).

3.1 Testimony and default entitlement

In Brandom's model, entitlement to a propositional commitment can arise in two ways: by inference from a commitment to which one is already entitled, or by deferral to the testimony of an interlocutor who is entitled to the commitment. Stated thus simply, there is an obvious threat of infinite regress on both scores, since it appears we may not acquire any entitlements unless there are already commitments that we or our interlocutors are entitled to. Brandom finesses this danger by proposing a "default and challenge" model: entitlement to a commitment is often attributed by default, though remaining potentially liable to be challenged by the assertion of an incompatible commitment Which commitments are taken to be prima facie entitled and which are liable to vindication is a matter of "social practice", though a little reflection will show that we go through our days attributing default entitlement to a great deal, perhaps most of the propositional commitments we encounter.

Brandom seems to have in mind relatively banal claims which it would be silly to challenge, such as "There have been black dogs" or "I have ten fingers". However I think we can safely go further than this, and assume that people are generally disposed to accept novel claims that do not conflict with their prior beliefs. [1] observe that human societies are characterised by "generally honest communication" and that humans tend to be "credulous": while this may leave us potentially vulnerable to free-riders such as gossips and rumourmongers, it is the price we have paid in cultural evolution for mostly stable societies and the rapid transmission of new ideas and novel practices. Crucially, Brandom claims that practical commitments are not transferrable in the same way: while performing an action incurs a commitment to justify it, it does not authorise others to carry out the same action.

Brandom's account of practical reasoning has received relatively little critical attention, by comparison with the account of propositional reasoning: it is explicitly excluded from a recent monograph on Brandom's philosophy [24] and none of the papers collected in [25] make it their focus. In fact I am not aware that the central claim of asymmetry between the two modes of reasoning has been challenged in Brandom commentary.

Brandom's account of action and intention is initially quite similar to his propositional story in its overall structure: the role of intentions is taken by practical commitments which can stand in inferential relations to propositional or other practical commitments, and to which one may be entitled or not entitled. It is notable that practical commitments can be inferred from propositional commitments as in examples like:

1. Only opening my umbrella will keep me dry, so I shall open my umbrella.
2. I am a bank employee going to work, so I shall wear a tie.

Brandom argues that these inferences are not enthymematic, relying on suppressed premises "I wish to stay dry" or "Bank employees should wear ties", but that (1) and (2) are in fact examples of what he (following Sellars) calls "material inference": the consequent follows from the antecedent by virtue of its content, and the putative "suppressed premises" are ways of making explicit the implicit norms or preferences that make the inferences go through.

Many people encountering Brandom's work find the notion of material inference puzzling and suspicious, particularly in the way it seems to provide free inference tickets for deriving "ought" from "is". Space does not permit an in-depth discussion of this issue: for now we merely note that practical commitments are taken to stand in inferential relations with both propositional and other practical commitments, and that an action is taken to be rational if it fulfils a practical commitment for which the agent can give a reason. For example: "Why are you wearing a tie?" "I'm on the way to work". Putting things a little more technically: to demonstrate entitlement is to offer a chain of reasoning which terminates in a practical commitment which is compatible with one's other acknowledged commitments, and actions result from "reliable dispositions to respond differentially to the acknowledgement of certain sorts of commitments" [3]. Scorekeepers are licensed to infer agents' beliefs from their intentional actions [Ibid.].

3.2 Commitment updates

Following [13] we assume that in a multi-agent interaction, each agent A_n maintains a "deontic scoreboard" for each agent A_i including sets C_i and E_i of commitments and entitlements which A_n attributes to A_i (including the case where $n = i$). Commitments will be stored as labelled formulae $L : \phi$ where ϕ represents a proposition and L details A_i's grounds for commitment or entitlement to ϕ (cf [6]. Update operations involve the following consequence relations:

$\Rightarrow_C$ committive entailment: commitment to P involves commitment to Q

$\Rightarrow_P$ permissive entailment: entitlement to P involves entitlement to Q

$\Rightarrow_\perp$ incompatibility: commitment to P precludes entitlement to Q

Various proposals have been made for the formal semantics of these relations. [15] proposes that committive entailment should be formalised using relevance logic while permissive entailment corresponds to classical logic, while [4] sets out a detailed semantic framework based on a fundamental notion of *incompatibility* and [17] proposes a natural deduction-based account of dialogue structure "in the spirit of Brandom's logical expressivism".

Labels on formulae may involve these relations to indicate the source of a commitment: L may present a proof of ϕ e.g.

$$\{\psi, \psi \Rightarrow_C \phi\} : \phi$$

or cite an external source of information, where A_j denotes a human or artificial informant:

$$defer(A_j, \phi) : \phi$$

or rely on a non-inferential belief derived from observation:

$$\{observe(A_n, \sigma), observe(A_n, \sigma) \Rightarrow_C \phi\} : \phi$$

or involve an abductive inference:

$$\{done(A_i, \alpha), \phi \Rightarrow_C \alpha\} : \phi$$

where α denotes an action carried out by A, and ϕ is a hypothesized reason for A to do this.

It is also assumed that each agent A_n has a private knowledge base of auxiliary hypotheses/beliefs, referred to as Γ_n, which is employed in calculating other agents' consequential commitments and entitlements. Assertions in Γ_n will also be labelled formulas annotated with a record of the source of information. So an assertion of ϕ by agent A_i or an action by A_i which presupposes commitment to A_i results in the following updates of A_n's information state [24]:

1. $C_i = C_i \cup \{\emptyset : \phi\}$ - add ϕ to A_i's commitments
2. $C_i = C_i \cup Cl(\{\{\Phi \wedge \phi \Rightarrow_C \psi, \phi\} : \psi \mid (\Gamma_n \cup C_i \Rightarrow_C \Phi) \wedge ((\Phi \wedge \phi \Rightarrow_C \psi)\})$ where Φ is an atomic or complex formula: add all committive consequences of ϕ along with existing commitments C_i and the scorekeeper's background commitments Γ_n.
3. $E_i = E_i - \{L : \psi | \exists L'\rho \in C_i : L'\rho \Rightarrow_\perp L : \psi\}$ - remove all commitments from the entitlement set which are incompatible with the updated C_i
4. $E_i = Cl(E_i)$ under $\Rightarrow_C$ - add all committive entailments of contracted entitlement set
5. $E_i = E_i \cup \{\emptyset : \phi\} \cup \bigvee(\{\{\Phi \wedge \phi \Rightarrow_P \psi, \phi\} : \psi \mid (\Gamma_n \cup E_i \Rightarrow_P \Phi) \wedge (\Phi \wedge \phi \Rightarrow_P \psi) \wedge (\neg\exists\Psi : C_i \Rightarrow_C \Psi \wedge \Psi \Rightarrow_\perp \psi)\})$ - add ϕ to the entitlement set along with the disjunction of all permissive entailments of ϕ and Γ_n - which need not be consistent with each other, but must all be consistent with the commitment set.
6. Finally: if ϕ is consistent with E_n, add $defer(A_i, \phi) : \phi$ to C_n and repeat 2 - 5 with n substituted for i. That is, if the scorekeeper A_n considers A_i is entitled to commit to ϕ, A_n can add ϕ to his or her own commitments and entitlements, with an indication that A_i is the source of the information.

3.3 Imitation within a rational practice

The aim of this and the next section is to show how conformist imitation can be modelled as part of a rational practice, involving agents who are capable of demanding and giving reasons for their actions. The use of labelled formulas to represent commitments is intended to facilitate this by encapsulating the inferential history and justification of individual commitments. In the event of disagreement, claims can be evaluated by comparing the reliability and trustworthiness of informants, strength of premises or the accuracy of a scorekeeper's hypotheses about the reasons for an action. So for example if A claims ϕ and B counter claims ψ s.t. $\psi \Rightarrow_\perp \phi$, A may then offer a justification $defer(C, \phi) : \phi$ which B counters with $defer(D, \chi) : \chi, \chi \Rightarrow_C \psi$, and the issue may be resolved by assessing whether C or D is considered a more reliable source.

4 PARALLELS BETWEEN PROPOSITIONAL AND PRACTICAL COMMITMENTS

As noted above, Brandom argues that there is a fundamental difference between the two flavours of commitment: there is "nothing corresponding to the authority of testimony in the practical case" [3]. That is: while "whatever is a good reason for one interlocutor to undertake a [propositional] commitment is a good reason for another

as well" [24] it is not generally the case that a good reason for you to perform an action is a good reason for me. Brandom gives as an example that he may have good reason to drive to the airport this afternoon, but this doesn't mean that I do. We may "have quite different ends, subscribe to different values, occupy different social roles, be subject to different norms" [2].

This can be challenged in a number of ways. First of all, it is questionable just how portable propositional entitlements really are in the limit. Of course in the ideal case, if John can display a chain of reasoning which is grounded in commonly accepted objective truths and justifies his commitment, then Mary can help herself to this argument as an entitlement to her own commitment to P. However, Brandomian agents can't in general be assumed to be in this happy state, and in fact many commentators have argued that Brandom fails to provide a convincing account of objectivity [24, 8]. Entitlements will always be provisional and defeasible, and they will be more or less available to different agents according to their own auxiliary commitments.

In fact Brandom acknowledges that beliefs may differ among individuals as much as desires do, but insists that there is "an implicit norm of common belief that has no analog for desire" [2]. He further argues that there is a fundamental difference between the practical and cognitive structures in that desires are a different class of entity from beliefs: the latter are propositional, functioning as premises and conclusions of inferences, while the former rather encode patterns of inference from doxastic to practical commitments (Ibid.). This is, as [10] notes, an unusual and counter-intuitive position, and is very much an artefact of Brandom's general model. In any case, this distinctive characterisation of desires need not preclude particular preferences being widely shared within a community - such as a wish to avoid getting wet, or to conform to general standards of attire.

Let us suppose that I have a settled opinion that other people are rational, in that they always have good reasons for what they do, and I further believe that if you have a good reason to do something, there may well be good reasons for me to do likewise - other things being equal, i.e. assuming I have no incompatible commitments. Of course you and I may operate according to different ends, values, norms and so on, but all of these could in principle be handled within the model by treating them as sources of commitments which lead us to follow different courses of action. So I might think, "yes, it would be a good idea to go to the airport if only I didn't have to give my lecture".

A second point is that while I may well observe that Brandom is off to the airport, he is not the only person in my purview: lots of people are doing lots of different things and I clearly can't copy them all. The key factor here is selective attention: just as we are not going to automatically believe (attribute default entitlement) to just anybody, nor are we going to habitually imitate just anybody [11].

Of course, going to the airport is a somewhat exotic example and the point may be easier to make with a more everyday scenario. Suppose I am visiting the University of Pittsburgh and after lunch, I see Brandom taking his tray to a particular trolley at the end of the cafeteria. I may well do the same thing and if asked why, it would be quite reasonable to say "Well, *he* did it". It is something of a truism that when we are in unfamiliar situations, we often model our behaviour on those we judge to be well-used to local customs. To revisit Brandom's tie-wearing example: suppose I go to my first day at work in an open-necked shirt, but I notice that all the other male employees are wearing ties. If I then decide to put on a tie for my second day, I would justify this with the argument "Everyone else is wearing ties, so I shall wear one".

What this is leading towards is the idea that *example* can in and of itself be one among many possible sources of (defeasible) entitlement to take on a practical commitment with which one has no pre-existing incompatible commitments. Of course I am by no means trying to show that simple-minded imitation is always or even usually a rational course of action. The claim rather is that conformist imitation can be modelled within a normative framework that is characterised by mutual accountability, contrary to Brandom's claim that practical commitments are never heritable in the way propositional commitments can be ("there is no general (even defeasible) presumption of heritability" [2]). One could argue that it is precisely such a defeasible presumption of heritability of practical commitments that underpins the legal doctrines of precedent and analogy. Particularly in Common Law jurisdictions such as the US and UK, these doctrines provide that in appropriate circumstances, a court of law may justify its actions with reference to similar actions previously taken by a competent body [14]. However, to pursue such an argument would take us too far from the concerns of this paper.

A background assumption underlying this argument is that if I am to regard another's action as entitling for me, I must also regard him as entitled to it. This points up a second, silent asymmetry in the accounts of propositional and practical entitlement in [2]: there seems to be no notion of default entitlement to practical commitments corresponding to that for propositional commitments. "Entitlement" here would mean that an agent can offer a chain of practical reasoning which begins with a propositional commitment (to which they are judged to be entitled) and ends with a practical commitment to perform the action in question. I suggest it is quite intelligible to propose that we habitually assume such an argument exists - that people have reasons for what they do - even if we are not in a position to reconstruct it with confidence.

Types of inherited entitlement

The following list gives details of some ways in which one may defeasibly claim inherited entitlement to an assertion or a course of action. Item (1) corresponds to Brandom's account of inherited entitlement to propositional claims, while (2) appears to be entirely consistent with his account, while it does not require that the material inference mentioned in (2b) is shared with other agents. (3) is the only mechanism Brandom seems to allow for transfer of practical commitments from one agent to another [2]. The remaining items illustrate further proposed extensions into the domain of practical reasoning: (4) can, I claim, be handled within the formalism sketched in section 3.2 above, though it does assume that the inference referred to in (4b) expresses a preference that is shared between myself and J. (5) is more speculative and more work would be needed to handle it within the formal system.

1. Testimony: I am entitled to assert P because J is committed to P, and I attribute to him default entitlement
2. I am entitled to do A because
   (a) I am entitled by testimony to assert Q
   (b) Commitment to do A follows from Q by material inference
3. I am committed and entitled to do A because
   (a) I am in a subordinate relation to J
   (b) J issues an order which imposes a commitment on me to make a particular assertion true.
4. I am entitled to do A because
   (a) I observe J doing A

(b) I infer that J is committed to the proposition P, and that commitment to do A follows from P by material inference

(c) I inherit entitlement to P by (inferred/ostensive) testimony

(d) I become committed and entitled to do A

5. I am entitled to do A because

(a) I observe J doing A

(b) I attribute to J default entitlement to do A

(c) I have no commitments incompatible with doing A

(d) I inherit entitlement to do A from J by example

5 DISCUSSION

If we interpret normativity in terms of mutual accountability, following e.g. [10, 18, 2], 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 claim and put into question entitlements to commitments. 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 [22].) We can however identify stages in the process where imitation as inheritance of practical entitlement would slot in. The transmission process is divided into on the one hand, vertical (parent-offspring) and horizontal (peer-peer) dimensions, and on the other active transmission, where norms are purposefully communicated by an agent typically accompanied by sanctions, and passive transmission or "social learning" where agents acquire norms by observing their neighbours and copying the behaviour of the more successful ones. The account of conformist imitation presented in this paper could be modelled as passive transmission along either the vertical or horizontal dimensions.

[5] introduces a further perspective on normativity: while norms may operate through a process of mutual accountability, the identities of agents who are deemed to be "worthy of representation and recognition" in this process is itself normatively shaped on the basis of such factors as gender, religion, citizenship and so on. Regrettably, her highly nuanced discussion of these notions is vitiated by a re-emergence of regulism in the idea that normative frameworks are orchestrated by "state power", as if the state were itself a monolithic entity with a common purpose.

6 Conclusion

We have argued that certain representative studies of normative agency using agent-based simulations are flawed in that they ignore the dimension of mutual accountability, which has been extensively discussed in the relevant philosophical literature. We have also proposed that one such philosophical account can be fruitfully extended to provide a framework for modelling social learning via conformist imitation. This framework is however still remote from any practical applications and the next research efforts will aim to further formalise and operationalise the framework drawing on existing work in agent-based modelling, argumentative agents and social simulation.

REFERENCES

[1] R. Boyd and P. Richerson, 'Culture and the evolution of human cooperation', *Philosophical Transactions of the Royal Society*, **364**, 3281–3288, (2009).

[2] R. Brandom, *Making It Explicit: Reasoning, Representing, and Discursive Commitment*, Harvard University Press, Cambridge, MA, 1994.

[3] R. Brandom, *Articulating Reasons: An Introduction to Inferentialism*, Harvard University Press, Cambridge, MA, 2000.

[4] R. Brandom, *Between Saying and Doing: Towards an Analytic Pragmatism*, Oxford University Press, Oxford, 2008.

[5] J. Butler, 'Non-thinking in the name of the normative', in *Frames of War: When is Life Grievable?*, 137–164, (2010).

[6] C. Chesñevar and G. Simari, 'Towards computational models of natural argument using labelled deductive systems', in *Proc. of the 5th Intl.. Workshop on Computational Models of Natural Argument (CMNA 2005), 19th Intl. Joint Conf. in Artificial Intelligence (IJCAI 2005). Edimburgh, UK*, eds., C. Reed, F.Grasso, and R.Kibble, pp. 32–39, (2005).

[7] R. Conte and F. Dignum, 'From social monitoring to normative influence', *Journal of Artificial Societies and Social Simulation*, **4**, (2001). http://jasss.soc.surrey.ac.uk/4/2/7.html.

[8] B. Hale and C. Wright, 'Assertibilist truth and objective content: Still inexplicit?', in *Reading Brandom: On Making It Explicit*, 276–293, (2010).

[9] C Hamblin, *Fallacies*, Methuen, London, 1970.

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

[11] J. Henrich, R. Boyd, and PJ. Richerson, 'Five misunderstandings about cultural evolution', *Human Nature*, **19**, 119–137, (2008).

[12] C.D. Hollander and A.S. Wu, '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] R. Kibble, 'Speech acts, commitment and multi-agent communication', *Computational and Mathematical Organization Theory*, **12**, (2006).

[14] G. Lamond. Precedent and analogy in legal reasoning. E. N. Zalta, (ed), The Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/archives/fall2008/entries/legal-reas-prec/, 2008.

[15] M. Lance, 'Two concepts of entailment', *Journal of Philosophical Research*, **XX**, (2006).

[16] M. Neumann, 'Norm internalisation in human and artificial intelligence', *Journal of Artificial Societies and Social Simulation*, **13**, (2010). http://jasss.soc.surrey.ac.uk/13/1/12.html.

[17] P. Piwek, 'Dialogue structure and logical expressivism', *Synthese*, **183**, (2011).

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

[19] B.T.R. Savarimuthu, S. Cranefield, M.A. Purvis, and M.K. Purvis, 'Obligation norm identification in agent societies', *Journal of Artificial Societies and Social Simulation*, **13**, (2010). http://jasss.soc.surrey.ac.uk/13/4/3.html.

[20] S. Sen and S. Airiau, 'Emergence of norms through social learning', in *Procs of IJCAI07*, pp. 1507 – 1512, (2007).

[21] M.P. Singh, 'A social semantics for agent communication languages', in *Issues in Agent Communication*, pp. 31–45, (2000).

[22] F. Toni, 'Argumentative agents', in *Procs of IMCSIT*, pp. 223–229. IEEE, (2010).

[23] D.N. Walton and E.C.W. Krabbe, *Commitment in dialogue: basic concepts of interpersonal reasoning*, SUNY series in logic and language, State University of New York Press, 1995.

[24] J. Wanderer, *Robert Brandom*, Philosophy Now, Acumen Publishing, 2008.

[25] B. Weiss and J. Wanderer, *Reading Brandom: on Making It Explicit*, Routledge, 2009.

Reputation Diffusion Simulation for Avoiding Privacy Violation

David Pergament[12], Armen Aghasaryan[1] and Jean-Gabriel Ganascia[2]

Abstract. When people expose their private life in online social networks (OSN), this doesn't mean that they don't care about their privacy, but they lack tools to evaluate the risk and to protect their data. To help them, we have previously designed a system called FORPS for Friends Oriented Reputation Privacy Score which evaluates the dangerousness of people who want to become our friends, by computing their propensity to propagate sensitive information. To anticipate the long-term and large scale effects of our system, we have built a multi-agent simulation that models a high number of interactions between people. We show that privacy protection based on different variants of the FORPS system produces better results than a simple decision process, in term of evaluation of the requestor's dangerousness, of convergence speed and of resistance to rumor phenomena.

1. INTRODUCTION

Numerous societal and ethical issues are related to the development of online social networks (OSN). Among them, the risks for the privacy protection have often been mentioned. On the one hand, some people are afraid of the risk that the individual data, such as photos or commentaries, become public or that the owners of the social network infrastructure exploit them for their own purposes without taking care of individuals' rights on those data. On the other hand, the social networks reshape continuously their privacy policy, taking into account the addressed criticisms and making people able to define by themselves the degree of visibility of their data.

To clarify the debate, let us remind that the privacy protection is based on a general principle according to which everyone has the right to totally control his personal information, i.e. to decide what information he/she accepts to reveal, when and to whom he/she does it [15]. However, this general principle is difficult to apply on social networks, because of the difficulty for a user to know who the persons asking him to be his 'friends' are and how they usually behave with their already existing friends.

In addition, individuals change with time and age. It may then appear necessary for them to hide photographies, movies or textual content that corresponded to part of their previous life. It corresponds to the notion of "right to forget", which means that individuals should be able to delete all the personal data they want. However, if we don't pay enough attention, social networks may contain huge quantities of individual data that can't be erased, especially if their supposed friends have divulged these data without asking their consent.

For all these reason, it appears necessary to help the individuals to define their privacy policy on social networks by warning them about the potential dangers of individuals they don't know, but who asked them to become their friends.

This is exactly what motivated the design of the FORPS ("Friends oriented Reputation Privacy Score"). Namely, we have we have introduced this system to control the propagation of information through social networks by scoring the propensity of individuals to propagate private information [10] [13]. Now, it's time to evaluate the effect of such a scoring mechanism on the actual propagation. We address this problem from two perspectives:

1. On one hand, we evaluate the legitimacy of the use of FORPS, and its efficiency in terms of convergence to a state where people have a correct a priori knowledge on a given requestor.
2. On other hand, we add dynamicity to our system: what happens if the requestor changes? What happens if malicious individuals try to propagate rumors?

By creating a high number of interactions with a simulator, our goal is to validate, anticipate and calibrate the properties of FORPS in such a way that they ensure its privacy goals at best, without acting against the requestor.

The paper is organized as follows. The second part refers to the related art of reputation scores and diffusion models. The third part presents the model used in the simulation. The first results are then presented in the fourth part. And finally we discuss the limits and the perspectives in the sixth part.

2. STATE OF THE ART

The technology presented in this paper is related to people scoring for which several works have been carried out.

In the domain of e-reputation, we can mention websites such as *www.123people.com* which find and aggregate data from different sources on the web and which provide information about an individual. Some systems, like Klout [3], measure the popularity of people, how much for example their action influence the others. We also have eBay's mechanism, where users can give notes about the degree of trust they have on somebody they dealt with before. Also, there are scores like the fico score [4] used to estimate the likelihood that a person will default on a loan. However, these systems are not really tackling privacy issues.

More related to privacy, we can mention various systems that have already proposed the concept of privacy score which can be used to alert users about the visibility and protection of their sensitive data. They are implemented as websites (e.g., Profile Watch [5]) or as Facebook applications (e.g., 'Privacy Check' [6]). Liu and Terzi proposed a privacy score on social networking sites. The scores are computed by considering two factors, the visibility and the sensitivity of the user's data [7]. Our privacy reputation score differs from the aforementioned approaches in that it takes different input data and uses a different algorithmic approach for the score computation [10]. Instead of analyzing

[1] Alcatel Lucent Bell Labs, Nozay 91620, France. {**david.pergament**, **armen.aghasaryan**}**@alcatel.lucent.com**.

[2] LIP6 – Université Pierre et Marie Curie, Paris 75252, France. **jean-gabriel.ganascia@lip6.fr**.

[3] **http://www.klout.com**
[4] **http://www.myfico.com/**
[5] **http://www.profilewatch.org/**
[6] **http://www.rabidgremlin.com/fbprivacy/**

only the data owner's private or public data, our approach also considers the particular usage context defined by another user (the data requester) who is requesting an access to the data owner's information. This request can be formulated either as a friendship request in a social network or any other request to access a specific content item of the data owner [13]. The score represents the estimated privacy risk to the data owner if the request is granted. We notice that [8] and [6] also point out that sensitive information exposure can be cause by your friends. But for the latest, they are dealing with global profile information (like age). Unlike FORPS, they do not take into account the textual contents.

Also, multi-agent simulators have been broadly used to simulate the diffusion process over real or online social networks. We are quite close to classical diffusion phenomena, provided we consider the diffusion of the requestor's score as for example, the diffusion of an innovation [12]. The authors of [1] have worked on privacy diffusion. But their goal was totally different; they wanted to simulate the migration of people from Myspace to Facebook for privacy reason.

3. FORPS: FRIENDS ORIENTED REPUTATION PRIVACY SCORE

The basic idea of the FORPS mechanism consists in taking advantage of the overall knowledge present in a social network and that is accessible to a given user (e.g., Alice). Then, the system tries to estimate the danger that another user (e.g., Calvin) may represent with respect to a non-desirable propagation of Alice's sensitive data. This can be done by aggregating different sources of information characterizing Calvin's profile and behavior:

1. public profiles of other users available in the social network or any public data on the web,
2. the private profile of Calvin, insofar as it is visible to the Alice, and more importantly,
3. the information that the friends of Alice possesses or have access to, concerning Calvin, such as likes or comments that Calvin leaves on photos belonging to one of the friends of Alice;

The FORPS system allows Alice to define her privacy sensitivity profile which is characterized by the themes/categories, the object-types that are relevant for Alice. For instance, Alice may want only some of her content items concerning a specific topic (e.g. family) to stay in a restricted area of users, other topics can be propagated. The same applies to different object types such as posts, photos, videos, etc. These preferences are taken into account by the system to calculate different privacy reputation scores of Calvin per theme and object type and then to obtain an aggregated score. Different semantic analysis techniques are used [11] to identify the appropriate themes for each user. The score computation is based on different behavioral factors characterizing information propagation in social networks, e.g. propagator propensity, information sensitivity, and user popularity. Some factors are quantitative; others are qualitative and pivot on sentiment mining analysis techniques [3]

By extension of the FORPS approach, in FORPS+ the scores are computed collaboratively: two users who have a high confidence relation (e.g., very good friends), can exchange their privacy score in order to combine their information about Calvin so that their computations became more accurate. This extension assumes that the scores have the same semantics for the two users. Namely, as the scores are theme-dependant, FORPS+ ensure the similarity of the sensitivity profiles.

4. SIMULATION MODEL

An online social network is modelled as an undirected graph G = (V,E) in which vertices (V) or nodes represent the individuals, and edges (E) represents a finite set of links between the individuals, usually a friend relationship, such that $E \subseteq V \times V$ (Mika, 2007). It can be represented by its symmetrical nxn characteristic matrix FS := $fs_{i,j}$, where n = $|V|$, and

$$fs_{i,j} = \begin{cases} 1 \,\|\, (vi, vj) \in E \\ 0 \,\|\, otherwise \end{cases} \quad (1)$$

The number of friends an individual have is called the degree of the corresponding node.

4.1 The Agents

Our simulation has four categories of agents:

1. The requestors. This category will be composed by only one agent, let's called him the agent 'r'.
2. The members 'c1' of the circle of 'r' in a primary social network. They are composed by the friends of 'r' as well as people 'r' wants them as friends (potential future friends), or wants to be aware of their activities ("subscribers" in Facebook, "circles" in Google+ [7]).
3. The members 'c2' of the circle of 'r' in a second social network. We simulate two different social networks in order to perform real-time comparisons. As we will see in the experimental part, these two social networks are twin. Each member of a social network has its alter-ego in the other.
4. The rumors launchers 'm' are users which trigger rumors regarding the requestor 'r'. Those agents have the faculty of not being influenced by other agents. They will propagate a message that is opposite to the true nature of 'r' (see the following chapter). We notice that this specific faculty can for example be possessed ex- friends, which have arrived to a point of no return regarding their negative confidence in 'r'.

4.2 The Diffusion Model

$S_c^t(r)$ represents the score of the requestor 'r' at a time 't 'according to 'c', a member of its circle. This score indicate the assumed degree of safeness of the requestor. The higher it is, the more 'c' consider 'r' as safe. The lower it is, the more 'c' consider 'r' as dangerous.

$S_r^t(r)$ represents the *real privacy score* of a requestor. As the requestor is the only entity that possesses all the information about him, we use the index 'r' (requestor) for this score. By considering that this value exists, we make here a strong assumption: we consider that the requestor has a coherent behavior at a given instance of time 't' which is moreover systematically reflected in its interactions with others users.

4.3 The Meetings

At each step (i.e. each iteration), agents move within the simulated 2D plane starting from original position and moving in randomly selected direction with a small step. When an agent 'c' is localized at the same position of 'r', there is a possibility that a direct or indirect information transfer occurs between 'r' and 'c' (see section 4). This communication event, Com(r,c), is triggered in the simulation model according to the following rule:

[7] As we are dealing with OSNs based on privacy, that's why we do not mention public OSN like Twitters, with its followers.

$$s(r,c) - |\theta_{com}| > s_{threshold} \quad (2)$$

where *s(r,c)* represents the strength of the friendship. This value depends on the presence of a friendship relation, $fs_{r,c}$, as well as on the number of friends in common between 'r' and 'c'. To trigger Com(r,c), *s(r,c)* is combined with a random perturbation θ_{com}, and checked against a system-wise defined threshold, $s_{threshold}$. We introduce a negative random perturbation to account for the situations where the information transfer is not meaningful with respect to the safeness degree of the requestor

As we want to give chances to a discussion to be continued, we need to give to our system a short-term memory. By reinforcing the probability of meetings that have already took place, the slight and random move policy fits well with this goal.

4.4 The Information Exchange

When a communication is happening, agents exchange information about the requestor. By saying "interacting", we have in mind the comment of a status, the 'like' a photo, the tag of an article etc… We have previously defined in FORPS [10] that in a social network context, exchanging information could be done directly (information accessible thru the own data of 'c'), or via a friend in common.

Let's suppose now that all the interactions that exist in our simulation are interactions between the requestor and the members of its circle, and that they can either represents a direct exchange or an exchange via common friends (indirect exchange). So, in FORPS, when an interaction occurs between 'c' and 'r' (i.e the communication event Com(r,c) is triggered), the new score (at t+1) of 'c' regarding 'r' will get closer than the real score of 'r' by being updated as follows:

$$S_c^{t+1}(r) := \alpha.S_c^t(r) + (1-\alpha).S_r^t(r) \quad (3)$$

In FORPS+, all the users that are in a friend relationship with the member who has interacted with 'r' will also benefit from the added information (provided that the addition is substantial: $S_c^{t+1}(r) - S_c^t(r) \geq \Delta$):

$$\forall c' / FS_{c,c'} = 1$$
$$S_{c'}^{t+1}(r) := \beta.S_{c'}^t(r) + (1-\beta)S_c^{t+1}(r) \quad (4)$$

where $\beta \leq \alpha$, indeed, the scores people have directly computed will have a higher impact because in this case, the requestor's data are analyzed with more personalized criteria [10].

The rumors launcher agents have the same power of a requestor: they can influence others (except the requestor himself). Mathematically, they behave as a requestor. When a member 'c' will meet a rumor launcher, (i.e this communication event Com(m,c) is triggered), it will increase the amount of information it has related to the requestor:

$$S_c^{t+1}(r) := \alpha_m.S_c^t(r) + (1-\alpha_m).S_m^t(r) \quad (5)$$

Note: We have considered here that the rumor is propagated within the score FORPS. This is a shortcut. We should better have an independent global opinion score, which would be composed by the FORPS score and the Rumor score. For our simulation, we consider here that the Rumor is an entry of the FORPS computation, even it is not a source of information created by the requestor himself as all the other entries.

4.5 The Instantiation of the Network

How can we simulate instantiations of real nodes of an online social network in our model?

Usually, the degree distribution (the distribution of friendship links in a network) follows a power-law distribution [14]. But in our case, as we focus on the requestor's networks, we don't consider the edges of all the nodes, except for the requestor's node. So we will just ensure the presence of simpler properties.

Iteratively, for each member of the circle, we choose randomly 3 others members, and we connect them with the first. We observe that with this simple algorithm, few members have a high level of connections, whereas the majority remains with a homogeneous number of connections.

4.6 Decision process and monitors

At the beginning, we have to define the real privacy score for the requestor: $S_r^t(r)$. Then, by interacting with the requestor (see section 4.4), the idea each member of its circle have on him (represented by $S_c^t(r)$) will change. Each member is unique: it has a personal acceptability threshold below which its opinion over the requestor becomes negative[8]. To simplify the simulation, we tolerate disloyalties: the possibility to a requestor's friend to often break its relation. And this is exactly what happened when its opinion become too negative: it breaks its relation with the requestor. Symmetrically, when its opinion become enough positive (relatively to its personal threshold) it re-establishes again its friendship relation. The figure 1 represents results obtained with our monitors:

1. The average opinion (Global Opinion monitor) on the requestor computed by all the members of its circle.
2. The number of requestor's friends in green (Friends monitor).
3. The number of people in its circle who are not its friends in red (Friends monitor).
4. The convergence (stable global opinion and stable number of friends) in this case is obtained after 6390 iterations, and as we can see in the figure 1, the global opinion is quite similar to the real privacy score of the requestor $S_r^t(r) = 67$.

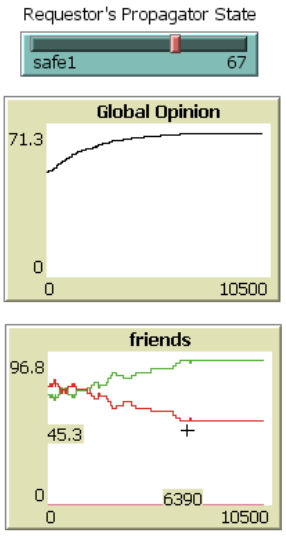


Figure 1: Monitors of our simulation

[8] In fact, the peer of alter ego within the two networks possess the same acceptability threshold

5. SIMULATION RESULTS

We decide to use the multi-agent programmable modeling environment NetLogo [16] in order to implement our models.

5.1 Preliminaries

1. Friendly Comparison Interface

We have designed a friendly interface which helps to compare the three models: Forps, Forps+, and "No". "No" is a simple model where friend's acceptance is only depending of the number of friends people have in common [2]. In fact, we were confronted initially to several difficulties linked to the simulation environment: from one experience to another, as the random parameters were different (especially the personal acceptability threshold and the links between agents) we were not capable to really compare two consecutive tests.

That's why we have implemented two parallel executions, with the same original parameters. The figure 2 shows how we can easily select the diffusion mechanism among the three that we propose. In the example of the figure 2, the social network 1 uses Forps as diffusion mechanism, whereas the social network 2 uses Forps+.

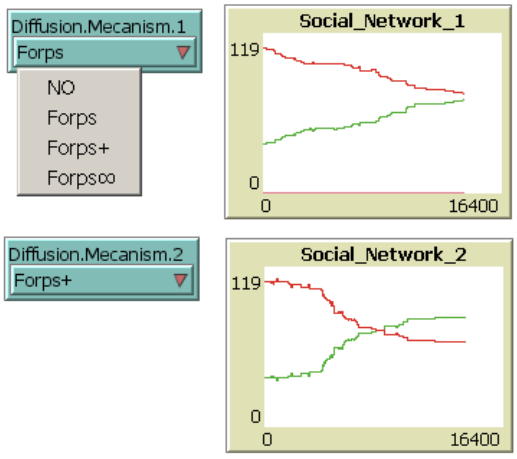


Figure 2: Diffusion mechanism selection

2. Comparison Indicators

- *Requestor' dangerousness evaluation error*. This is the most important indicator. It measures how far from the real score, the evaluation score is. It is the subtraction in absolute value of the two scores, the lower better.
- *Convergence speed*. During a simulation, agents are moving, and sometimes a communication occurs with a requestor or with a malicious agent (see section 4). Each of this communication steps is considered as an iteration. When the number of friends stops evolving, the simulation is over. The convergence speed index represents the number of iterations before the final convergence.
- *Half-life*. This is also an indicator of convergence. It informs when 50% of the agents in the circle of the requestor are its friends. If the requestor has a low score, half-life index may not exist. Note that this is also the intersection point of the red and the green curves where proportion of friends (in green) and non-friends (in red) is equal, see figure 1.

The three indicators represent the average value of the 300 simulations used in our experiment.

5.2 Forps' Legitimacy

1. With Forps, without Forps

We have conducted 300 tests for each of the models. For each test, the requestor has a fixed real dangerousness value and its circle is composed by 144 individuals. We have given the opportunity to the requestor to have 144 friends because this number is known in literature as the average number of friends of a Facebook user [5]. A test has duration of around 22 seconds.

The average results are represented in the following tables.

Comparison Indicators	Requestor's real score = 82		
	No FORPS	FORPS	FORPS +
Convergence	107,13	21422,38	16160,45
Half-life	82,72	12416,18	8508,88
Dangerousness evaluation error	20,20	1,11	0,96

Comparison Indicators	Requestor's real score = 55		
	No FORPS	FORPS	FORPS +
Convergence	102,55	16161,58	12814,87
Half-Life	18,26	No	No
Dangerousness evaluation error	45,12	1,20	1,03

We see that the convergence speed of a simple system (No FORPS) is the best. But it often gives absurd results (especially in the Table II) because it doesn't take into account the propensity of the requestor to propagate information. Indeed, the requestor may be dangerous, but because its circle members have more and more friends in common with the requestor, they accept gradually accept him as a friend.

2. Forps versus Forps+

We have implemented within NetLogo a way for triggering and analyzing a large quantity of tests, let's see deeper what happens when we compare FORPS and FORPS+. In the example below, the purple curve represents FORPS and the green curve represents FORPS+.

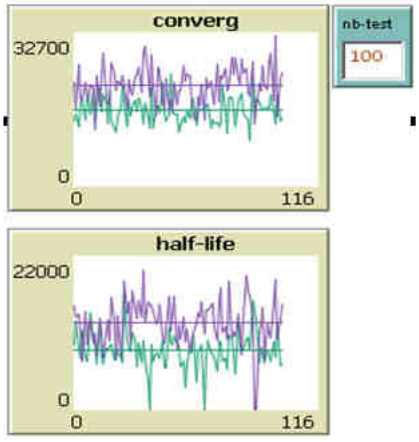


Figure 3: Indicators FORPS versus FORPS+

These plots show the series of values of three indicators taken from 100 tests. In terms of convergence speed, FORPS+ gave better results than FORPS at 86% of cases (see the figure 4). Note that it is not obvious to determine when a simulation has reached its stationary point (termination of the simulation). In

some cases, most of the agent will quickly reach their final states whereas some of them will conclude lately, because they have a selective acceptability threshold

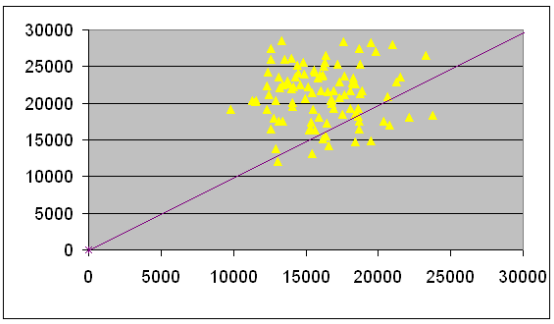


Figure 4: FORPS+ in x-axis, FORPS in y-axis

However, we can notice than if we exclude from the evaluation the simulations which have taken too much time (relatively to the others), FORPS+ become better than FORPS at 93% of cases.

5.3 Reactivity to requestor's change.

As we want to confer to our system a "right to forget" component, we want it to be able to amplify the impact of recent activities with respect to the old activities. Our logic is to catch the latest evolution of the character of the requestor. Let's see what its reactions in case of such evolutions are.

We can observe in the figure 5 that unlike the simple strategy, FORPS reacts quite well to this dynamics. Indeed, we see that the Global Opinion gradually becomes coherent with the requestor's real score. By conferring to our system such a property, we give clearly to the requestor the opportunity to give another opinion of him.

Note that this "right to forget" property of our system is different from a simple data aging over the time. In fact, if the requestor's real score remains unchanged, nothing would change in the score estimations neither.

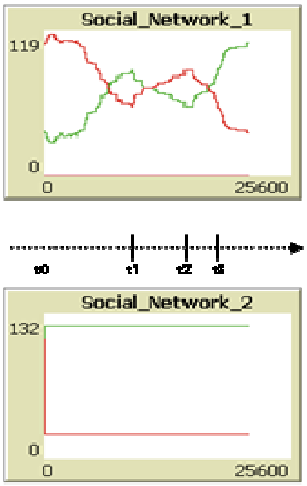


Figure 5: FORPS' reactions to requestor's change

t0: $S_r^{t0}(r)$**=81, t1:** $S_r^{t1}(r)$**=66, t2:** $S_r^{t2}(r)$**=83, t3:** $S_r^{t3}(r)$**=93.**

(Social Network 1: FORPS, Social Network 2: NO)

5.4 Reactivity to malicious rumors

An important property we want to obtain is the capability of the system to discern between authentic and false information available about the requestor. This is especially the case when a rumor appears. Let's have a simple scenario: a leader M and six of its active militants (AM(M)) want to explicitly propagate negative ideas on our requestor.

$$S^t(r)=70$$

$$S_m^t(r)=30 \ \forall m \in AM(M)$$

Let's see what will be the reaction of our simulated system on the figure 6.

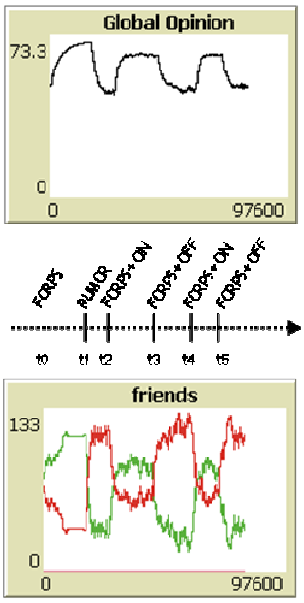


Figure 6: Reactions to malicious rumors

Before instance *t1*, the system has reached a stable state with FORPS option. At *t1*, the malicious rumor is triggered, its impact is radical. After only a few iterations, the requestor has inverted its proportion of friend (in green) and non-friend (in red) in its circle. The Global Opinion was also diminished but remained higher than 50% thanks to the effect of FORPS.

At the instance *t2*, we modify the propagation of the privacy score by applying FORPS+ instead of FORPS, and we observe that FORPS+ manage to contain the rumor. Indeed, the majority of the member of its circle becomes its friend again. This experience is repeated several times (t3, t4, t5 …) and equivalent results are obtained.

Contrarily to other phenomena considered in this paper, we do not observe a full convergence, but an oscillatory state, which is quite stable yet.

Finally, we then trigger the experiments with the two networks in parallel to have a synchronous comparison (Social Network 1: FORPS, Social Network 2: FORPS+)

We see in Figure 7 that when FORPS loose 20 points, between t1 and t1' (real score 80, Global Opinion 60) FORPS + loose only 11 points (Global Opinion 69).

We then modify the real requestor's score 80-> 92 (t2) during the same simulation. This can be considered as a reaction of counter-attack to the rumor from the point of view of the requestor. And we observe that FORPS+ manage to pass the half-life point, whereas FORPS does not.

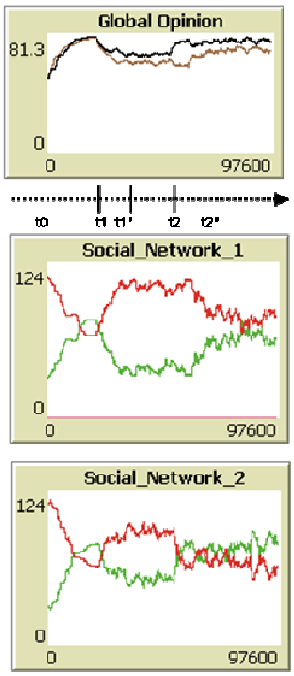


Figure 7: FORPS+ and FORPS' reactions to rumors

6. CONCLUSION AND PERSPECTIVES

The FORPS (Friends Oriented Reputation Privacy Score) system evaluates the dangerousness of people who want to become our friends, by computing their propensity to propagate sensitive information. In order to anticipate the long-term and large scale effects of this system, we have built a multi-agent simulation that models a high number of interactions between users. We have shown that privacy protection based on different variants of the FORPS system produces better results than a simple decision process, in term of evaluation of the requestor's dangerousness, of convergence speed and of resistance to rumor phenomena. Below we discuss several other findings and present the perspectives of the current work.

1. *Bootstrap Problem.* One of the assumed weaknesses of Forps was linked to the classical bootstrap problem [4] [10]. When we do not have any information about the requestor, how to initiate the process? What score should the system give for the requestor? In this paper, we have tested many initial states (very good, very bad, totally random, semi random). We find that the initial state has only a little influence on the convergence speed. Indeed, they all lead to the same final state which allows to conclude that bootstrap problem is not a problem for this system.

2. *The "NO" model.* Based on our intuitions and the previous work [2] we have supposed that in a simple process, people accept friend's requests when they have enough common friends with the requestors. We should include to this model a "loose friend" process. We plan to retrieve data related to the loss of friends over the time, by for example using tools as "unfriends"[9].

3. *Simple Simulation Model* One of the positive aspects of our simulation is that it is very simple. But we have not taken into consideration some aspects of the Forps process.

First, all the interactions we generate are considered faithful to the real privacy score of the requestor. But in real life, even if this score is quite bad, the requestor doesn't act every time negatively. He has also neutral or positive behaviors. For the moment, we have solved this problem by adding the random perturbation in the event triggering logic (see formula (1)). When it gives low number, it means that the discussion was not meaningful, and so it may be not considered as an interaction. The drawback is that this won't be considered as a positive interaction. The advantage is that it simplifies the simulation process. In a future work we intend to validate the assumption that such a simplified model does not perturb the final state of the estimated scores.

Second, in this simulation we have supposed that the focus is on a single topic. It has simplified our way to take into account the exchanges of scores between users (FORPS+). Everything was considered as meaningful because related to a sensitive topic for everybody. For further works we should introduce themes and give different sensitive profiles to the agents.

Third, we should also consider other specificities of the FORPS+ model. For example, we should favour the scores from friends who don't have exactly the same friends in common than me. Indeed, as their scores were computed by analyzing the same data, they won't really bring me new information.

4. *Testing with real users.* Finally, we envisage testing different variants of FORPS system within a corpus of real users in order to benefit from their feedbacks with respect to both usability aspects as well as the efficiency of different algorithmic parameters we have exploited in the simulated model. This will allow notably to validate the assumptions and the results derived from the simulations presented in the current paper.

[9] **http:// www.unfriendfinder.com/**

REFERENCES

[1] N. Baracaldo, C. Lopez, M. Anwar, M. Lewis. Simulating the effect of privacy concerns in online social networks. Information Reuse and Integration (IRI). IEEE International Conference on Digital Object Identifier (2011).

[2] Y. Boshmaf, I. Muslukhov, K.Beznosov, M. Ripeanu. The socialbot network: when bots socialize for fame and money. In Proceedings of the 27th Annual Computer Security Applications Conference (2011)

[3] A. Esuli, F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (2006)

[4] T. Gediminas, T. Alexander. Toward the next generation of recommender systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transaction on Knowledge and Data Engineering 17(6), pp.734—749 (2005)

[5] S.A. Golder, D.M. Wilkinson, and B.A. Huberman. Rhythms of social interaction: Messaging within a massive online network. 3rd International Conference on Communities and Technologies (2007).

[6] P. Gundecha, G. Barbier and H. Liu. Exploiting Vulnerability to Secure User Privacy on a Social Networking Site. In the 17th ACM SIGKDD (2011).

[7] K. Liu, and E. Terzi. A Framework for Computing the Privacy Scores of Users in Online Social Networks. In ACM Transactions on Knowledge Discovery from Data (2010)

[8] D. Massad. Herd Privacy: Modeling the Spillover Effects of Privacy Settings on Social Networking Sites. The Computational Social Science Society of the Americas (2011).

[9] P. Mika. Social Networks and the Semantic Web. volume 5 of Semantic Web and Beyond Computing for Human Experience (2007)

[10] D. Pergament, A. Aghasaryan, J. Ganascia, and S. Betge-Brezetz. FORPS: Friends-Oriented reputation privacy score, in Proceedings of ACM/IEEE International Workshop on Security and Privacy Preserving in e-Societies (2011)

[11] D. Ramage, D. Hall, R. Nallapati, C.D Manning. Labeled LDA: A supervised topic model for credit attribution in multi-label corpora. EMNLP (2009).

[12] E. Rogers. Diffusion of innovations. Glencloe (1962)

[13] Y. Wang, A. Aghasaryan, A. Shrihari, D. Pergament, G. B. Kamga, S. Betg e-Brezetz. Intelligent Reactive Access Control for Moving User Data. The Third IEEE International Conference on Information Privacy, Security, Risk and Trust (2011)

[14] S. Wasserman, K. Faust. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press (1994).
[15] A. Westin. Privacy and Freedom. Atheneum, New York. (1967).
[16] U. Wilensky. NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL (1999).

Understanding the formation and evolution of collaborative networks using a multi-actor climate program as example

Bei Wen[1,2] **and Edwin Horlings**[1]

Abstract. The mechanisms governing the composition of formal collaborative network remain poorly understood, owing to a restrictive focus on endogenous mechanisms to the exclusion of exogenous mechanisms. It is important to study how endogenous network structure and exogenous actor behaviour influence network formation and evolution over time. Current efforts in modelling longitudinal social networks are consistent with this view. The use of stochastic actor-based simulation models for the co-evolution of networks and behaviour allows the joint representation of endogenous and exogenous mechanisms, specifically the structural, componential, functional, and behavioural mechanisms of network formation. In this paper we study the emergence of collaborative networks in the Knowledge for Climate (KvK) research program. Endogenous mechanisms (transitivity and centrality) play a key role in the evolution of the KvK network. The results also reveal the influence of exogenous mechanisms: actors tend to collaborate with other actors from the same type of organizations (componential) and patterns of collaboration are affected by the nature and differences in roles (functional). Our analysis reveals a gap between actors from different sectors and a gap between actors working on global problems and those working on local problems. This is particularly visible in the fact that organizations active in hotspots projects, which focus on developing practical solutions for local and regional problems, are significantly more likely to form new ties than those active in theme projects.

1 INTRODUCTION

Networks have become a central concept in many fields, particularly in the areas of communication and organization. Among the various types of networks, collaborative networks are of special importance [1]. Collaborative networks are undergoing dramatic changes driven by scientific, economic, political, societal, cultural, and communicative processes collectively known as globalization [2].

These changes are particularly visible in science itself. In addition to the rise of international collaboration, scientific research is increasingly carried out in interinstitutional and international collaborative teams. Team science has evolved as a way to organize scientific research aimed at understanding and solving the most complex problems that confront humanity [3,4]. The rise of team science has created an urgent need to understand the fundamental configurations and interaction rules that govern the formation of collaborative networks as well as the behavioural patterns that emerge.

Understanding collaborative networks in science requires that we take into account two aspects of their evolution: complexity and history. Complexity arises from the fact that the actors in collaborative networks are largely autonomous, geographically distributed, and heterogeneous in terms of their operating environment, culture, social capital, and goals [1], have a set of attributes and preferences, and follow rules of interaction. They collaborate with each other to seek complementarities that allow them to participate in a competitive socioeconomic environment and achieve scientific excellence [5]. The history of networks relates to the fact that 'networks from nowhere' do not exist. Understanding the evolution of networks necessitates longitudinal analysis.

One way to analyse the formation of a complex social network is to simulate its emergence from the behaviour of individuals in the network. Simulation requires empirical data to verify the results.

We contribute to the understanding of the evolution of scientific networks and the empirical basis for future simulations by studying the Knowledge for Climate (KvK) research program, a €90 million multi-actor program aimed at developing useful knowledge for practical solutions to climate adaptation and mitigation.[3] Climate change is one of today's grand challenges and network effects are prevalent in climate science. The core of the program is formed by so-called hotspot projects in which government, industry, and science collaborate to develop real options for coping with climate issues at the local and regional level (e.g. in the port of Rotterdam and around Schiphol Airport).

The mechanisms underlying the processes of network evolution are not yet fully understood [6,7]. A deeper understanding of network evolution requires studying mechanisms that extend beyond the well-accepted drivers. The sociological literature on network formation and stability suggests four general mechanisms that may generate and sustain social ties that are potentially important for the KvK networks being studied, namely structural, componential, functional and behavioural mechanisms [8]. Our interest in both endogenous and exogenous mechanisms of network formation is linked with the recent theory on the co-evolution of social networks.

[1] Dept. of Science System Assessment, Rathenau Institute, 2593HW The Hague, The Netherlands. Email: {b.wen, e.horlings}@rathenau.nl.

[2] Dept. of Water Management, Faculty of Civil Engineering & Geosciences, Delft Univ. of Technology, 2628CN Delft, The Netherlands. Email: {B.Wen}@tudelft.nl.

[3] This paper was written as part of the project "Comparative Monitoring of Knowledge for Climate", which is carried out in the framework of the Dutch National Research Programme Knowledge for Climate (http://www.knowledgeforclimate.org).

The use of stochastic actor-based simulation models for the co-evolution of networks and behaviour allows the joint representation of endogenous and exogenous mechanisms and making the distinction between social selection and social influence processes, as elaborated by Snijders et al. [9,10,11,12]. Thus, we add to the empirical foundations of network simulation.

In section 2 we introduce the mechanisms of network formation and evolution. Section 3 describes the network data obtained from the KvK research program and outlines our approach to the analysis of structure, behaviour, and their dynamics. The results of the empirical study are presented and interpreted in section 4. Finally, in section 5, we present our conclusions and discuss our findings in light of the theoretical and practical relevance.

2 MECHANISMS OF NETWORK FORMATION AND EVOLUTION

The evolution of a network is driven simultaneously by endogenous effects that derive from network structure and actor positions, and exogenous effects that derive from the attributes and behaviours of individual actors. The combination of endogenous network effects and exogenous actor covariate effects constitutes the so-called objective function. This objective function captures the theoretically relevant information that the actor has at his disposal in the decision to establish a new tie or not [12].

Utilizing insights from the sociological literature on network formation, we have identified four general mechanisms that generate and sustain social ties that are potentially important for the KvK networks [8].

- *Structural mechanisms (endogenous).* The structural dimension addresses the structure or composition of the actors attached to the network. One of the principal features in most networks is the tendency toward transitivity or transitive closure. This means that collaborative partners of collaborative partners tend to become collaborative partners themselves. A second feature is that popular or active organizations will become even more popular or active in the collaborative network over time. Thirdly, The number of organizations with which an organization indirectly collaborates (i.e. the number of alters at geodesic distance two) is also considered to measure the effect from indirect relations. The tendency to keep other organizations at distance two can also be interpreted as negative measure of triadic closure.
- *Componential mechanisms (exogenous).* It has been argued that the identity of organizations constitutes an important aspect of form [13]. Individuals with the same type of affiliations tend to recognize each other's configurations of characteristic, processes, and resources [14]. The homophily principle, which suggests that collaborative partners are selected based on the similarity of characteristics, has been shown to be a crucial network mechanism in many contexts [15]. A second componential mechanism is geographic distance to the network centre and between individual nodes. The existing literature finds that geographical distance matters and that being geographically close stimulates and facilitates collaboration [16].
- *Functional mechanisms (exogenous).* This dimension considers the extent to which participants possess valuable and complementary competencies that help ensure the success of the collaboration [17]. Competencies represent the organization's knowledge, skills and capabilities. The individuals of the organizations active in the KvK program network play different roles, ranging from purely formal, non-substantive roles (e.g. legal representative, contract signee), programme functions (e.g. programme administrator, project supervisor), substantive roles in projects (e.g. project member, hotspot member), and leaders of projects, consortia, and hotspots. Theories of status variation address the greater capacity of high-status actors to attract others, compared with low-status actors [18,19].
- *Behaviour mechanisms (exogenous).* Behavioural approaches are based on the extent of participation behaviour at an organizational level. This contributes to our understanding of how the behaviours of individual organizations affect their chances of engaging in the collaborative network. It is proposed that organizations are more likely to engage in projects with established or experienced partners to maximize collective value.

Theories of network selection propose that the choice of network ties depends on the attributes and network embeddedness of actors as well as their possible alters. Social influence means that the behaviour (which also represents characteristics, attitudes, performance, etcetera) of actors depends on their own attributes and network position, but also on the attributes and behaviour of the actors with whom they are directly or indirectly tied in the network. In our paper, we presume that the relationship between participation and network formation may be explained by selection (ego seeks highly participating alters) or by influence (alters' participation influences the participation of ego). Each process has different implications. Determining the direction of causality is important for understanding the potential contribution of network dynamics [20].

Models have also been developed for the evolution of non-directed networks, such as collaboration networks, alliance networks, and knowledge sharing networks. For example, [21] studied the effect of job mobility of managers on inter-firm networks; [22] explained the development of interorganizational networks; [23] investigated the industrial alliance networks and found that reputation based on past performance was a strong predictor of alliance formation; and [24] examined how to facilitate innovation spreading in knowledge sharing networks.

3 DATA AND METHODS

The KvK research program is an ongoing collaborative program that was started in 2008. The program can be regarded as a constantly evolving social network of temporary collaborations [25,26]: collaboration is organized on the basis of projects that dissolve once the project, for which organizations are specifically set up, is completed. It includes 108 distinct but interrelated projects, and involves 102 organizations. The entire project and membership database of the KvK research program has been made available by the programme office. The master database has been cleaned and coded, and currently contains extensive information linking 1,131 individual members to projects, recording the starting and ending dates of their involvement in projects, showing the roles the individuals played

in projects and the organization the individuals represent, and indicating the theme to which the project belongs.

The data include details about the individual and institutional program members, the nature and timing of their involvement in different projects, as well as data describing the various projects. This allows us to examine how organizations and individuals collaborate and to study the mechanisms that facilitate or inhibit network formation and evolution.

Using this information, we constructed non-directed one-mode networks at an organizational level based on a binary association matrix indicating how individuals are indirectly linked with each other through the same project. This resulted in a symmetric association matrix of organizations with 102 rows and columns, where '1' represented a non-directed tie in which the row organization participated in the same project as the column organization, and '0' represented the absence of a tie.

The networks were divided into four waves according to the project periods: 2008, 2009, 2010, and 2011. The relationship between the organizations in each wave was visualized using Gephi [27]. The input information included (1) the association matrix, (2) the type of organizations, and (3) the geographic longitude and latitude coordinates of the organizations.

The similarity between consecutive waves was measured using the Jaccard index. The index is calculated as the number of ties present at both consecutive waves divided by the combined total number of ties. Since it is generally assumed that the change process is gradual, the Jaccard value should preferably be higher than 0.3 [12].

We use RSIENA to conduct stochastic actor-based simulation as described in [9], [10], [11], and [12] to estimate and evaluate a set of parameter values of interdependencies specified in an objective function that describes the development of KvK networks.[4] One advantage of RSIENA is that it allows us to infer the direction of causation between network selection and social influence [11,20]. Stochastic actor-based simulation has proved highly suitable for analysing longitudinal social network data and was specifically designed for estimating actor-driven network dynamics.

The set of parameters, or independent variables, include items that capture the structural, componential, functional and behavioural mechanisms, as described in Table 1. These parameters were first tested by score-type tests for statistical evidence about their effects without controlling for the effect on each other. The significant parameters were selected as the best specification for simulations.

Algorithmically, the simulation procedure begins with a set of preliminary estimates of the parameters, iteratively producing a sequence of parameter estimates based on a continuous-time Markov process, then comparing the resulting network and attribute matrices with the observed network data, and updating parameter values to reduce discrepancies. These iterative processes are repeated until the deviation between the parameter values and predetermined target values (t-ratio) are smaller than 0.1. The final parameter estimates are then used to simulate a new set of networks. In the simulations, we derived the standard errors of estimation for each parameter based on the set of simulated networks [9]. We constructed rate parameter models to assess the amount of change between consecutive waves, i.e. the speed with which the dependent variable changed. Three set of simulations were done, based on different models. The baseline model (model 1) included the set of significant parameters verified by score-type tests. The baseline model was then extended to incorporate both selection and influence processes. The organizational participation behaviour for the network and behaviour dynamics was tested in model 2. In model 3, we added control variables to balance the effects across groups.

Finally we used a function in RSIENA to assess the fit of model with respect to auxiliary functions of networks. The auxiliary functions concern the attributes of the network, such as degree distributions, which are not included among the target statistics for the effects in fitted models. Goodness-of-fit was visualized using "violin plots". A p-value for the goodness-of-fit was derived from a Monte Carlo Mahalanobis Distance Test [28]. The null hypothesis for this p-value is that the auxiliary statistics for the observed data are distributed according to the distribution simulated in phases of the estimations.

Parameter	Description or definition
Structural dimensions (endogenous)	
Degree (density)	(Intercept) Representation of the tendency to connect with arbitrary ties. Normally it is a negative value indicating the unlikelihood of forming ties randomly.
Transitive triads	Defined by the number of transitive alters in one ego's relations.
Degree popularity	Defined by the the sum of square root of the degree of the alters.
Indirect relations at distance 2	Defined by the number of alters at geodesic distance two.
Componential dimensions (exogenous)	
Identity	Defined by the type of organizations (program center, university, other knowledge institutes, government, firms, and NGOs and knowledge platforms).
Geodistance	Calculated by the logarithm of the geographical distance from each organization to the program center.
Geoproximity	Calculated by the logarithm of the geographical distance between each two of organizations.
Functional dimensions (exogenous)	
Role_max	Calculated by the highest role among individuals of each organization.
Role_average	Calculated by the average role among individuals of each organization.
Behavioral dimensions (exogenous)	
Role_sum	Calculated by the sum of roles of individuals belonging to each organization.
Individual_sum	Calculated by the number of individuals belonging to each organization.

Table 1. The description of dependent variables.

4 RESULTS

Figure 1 and Table 2 present the basic properties of the KvK network over time. They show how the network experienced a boost at the beginning and moderate changes in the following years. Over time, the network became more dense (graph density) and the number of collaborative partners of organisations increased (average degree). The changes of ties in consecutive networks, shown in Figure 1, were treated as the dependent variable in RSIENA modelling.

RSIENA program needs a certain amount of variation in ties between the network waves to be able to estimate the parameters. Jaccard coefficients for the similarity of consecutive networks were 0.140, 0.582, and 0.791, indicating an increasing

[4] The R software package RSIENA is freely available at http://www.stats.ox.ac.uk/~snijders/siena/siena.html.

similarity between the four waves. The Jaccard coefficients suggest that waves 2, 3 and 4 are best suited for modelling, because the change processes became gradual after wave 1.

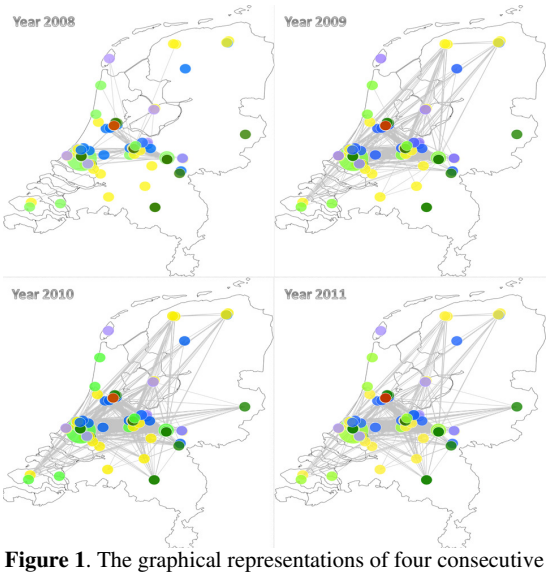


Figure 1. The graphical representations of four consecutive snapshots of KvK collaboration networks from 2008 to 2011. The nodes represent the organizations located geographically on a map of the Netherlands. The colour of nodes indicates the identity of the participating organizations, namely 3 program centres (red), 29 universities (dark green), 17 other knowledge institutes (light green), 28 government (yellow), 17 industrial firms (blue), and 8 NGOs or other knowledge platforms (purple). The existence of a collaboration tie between a pair of organizations is indicated using a solid grey line linking two nodes.

Observation time	Wave 1 (2008)	Wave 2 (2009)	Wave 3 (2010)	Wave 4 (2011)
Graph density	0.023	0.121	0.202	0.160
Average degree	2.294	12.196	20.431	16.157
Number of ties	117	622	1042	824

Table 2. Network density indicators

The modelling results are presented in Table 3. We began the analysis by simulating the endogenous and exogenous mechanisms. Model 1 in Table 3 shows all 12 identified parameters postulated for KvK network change and stability, including considerations of structural, componential, functional and behavioural dimensions. They were statistically verified with an acceptable fit to the data.

Structural parameters have a pronounced effect on network evolution. First, the negative effect of density (beta = -3.16, P < 0.001) is consistent with established knowledge obtained for most sparse networks [12]. This negative effect can be interpreted as an intercept, indicating that the costs of forming an arbitrary tie outweigh the benefits. In our case this suggests that it is unlikely that organizations form ties randomly. Second, KvK networks tend to be closed or transitive, as seen in the significant effects of transitive triads (beta = 0.48, P < 0.001). This finding is consistent with previous literature stating that collaborative partners of collaborative partners tend to become collaborative partners. Degree popularity (the square root of the degree of alters) measures the extent to which organizations tend to seek or be sought in the collaborative network. The positive effect size (beta = 0.47, P < 0.001) suggests that central organizations in the KvK network become even more central over time. The benefit of forming a tie must compensate for the cost per tie. Our results suggest that organizations should collaborate with a very central organisation with at least 45 relations in order to compensate for the -3.16 cost of creating a new collaboration ($0.47*\sqrt{45} = 3.16$).

Componential mechanisms involve the identity of collaborating organisations. There is a significant segregation according to identity (beta = -0.37, P < 0.001), meaning collaboration in the KvK program is influenced by the organization type. Moreover, organizations tend to collaborate with the same type of organizations (beta = 0.65, P < 0.001).

To measure the *functional* mechanisms, we weighted actor roles according to the substantive nature of their involvement in projects. The negative parameter estimates (beta = -0.44, P < 0.001; beta = -0.68, P < 0.001) imply that the more concrete the role actors played, the less likely it was that they sought for more network ties. For example, project leaders or principal investigators (weighted higher) appear less likely to connect to others, compared with regular project members (weighted lower). In addition, actors were less likely to participate in relations with actors having the same roles (beta = -3.03, P < 0.001). This effect may reflect a task division within collaborative projects, in which organizations jointly participated with a diversity of roles.

We found no significant effects among the *behavioural* mechanisms. Model 2 also incorporates the dynamics of behaviour, which models the organizational behavioural changes as a function of itself and the network evolution. The results showed that past participation behaviour had a significant effect in the long run (-0.06*(the extent of participation) + 0.00*(the extent of participation)^2). The average of alters' behaviour also had a significant influence on the ego's participation behaviour (beta = 0.00, P = 0.046), which means that organizations tend to adapt their participation behaviour to the average behaviour of their collaboration partners. However, all these effects are very small. Therefore, the evidence for participation-based social influence is weak.

The KvK research programme consists of eight geographical hotspots (Schiphol Mainport, Haaglanden Region, Rotterdam Region, Major rivers, South-West Netherlands Delta, Shallow waters and peat meadow areas, Dry rural areas, Wadden Sea) and eight research themes (climate proof flood risk management, climate proof fresh water supply, climate adaptation for rural areas, climate proof cities, infrastructure and networks, high-quality climate projections, governance of adaptation, decision support tools). Hotspot projects are the essence of the program. They were developed around specific locations in the Netherlands which are particularly vulnerable to the consequences of climate change. These locations function as real-life laboratories where knowledge is put in practice. Given the special functional and geographical importance of hotspot projects, we have tested the effects of project type (hotspots or not) separately in Model 3.

Table 3. Parameter estimates of KvK evolution model, with standard errors and two-sided p-values.

Effect	Model 1 (Baseline Model)			Model 2 (Bahaviour Dynamics)			Model 3 (Control Variable)		
	Estimates	SE	p-value	Estimates	SE	p-value	Estimates	SE	p-value
Network Dynamics:									
Rate function:									
0.1 Network rate period 1	4.65	0.23		4.61	0.27		4.92	0.26	
0.2 Network rate period 2	5.16	0.41		5.65	1.17		5.02	0.38	
Objective function:									
Structural dimensions (endogenous)									
1. Degree (density)	-3.16	0.40	0.000 ***	-2.44	0.09	0.000 ***	-3.20	0.35	0.000 ***
2. Transitive triads	0.38	0.06	0.000 ***	0.41	0.06	0.000 ***	0.36	0.04	0.000 ***
3. Degree popularity	0.47	0.11	0.000 ***	0.27	0.07	0.000 ***	0.44	0.11	0.000 ***
4. Indirect relations at distance 2	-0.05	0.04	0.206	-0.03	0.03	0.333	-0.06	0.04	0.069 +
Componential dimensions (exogenous)									
5. Identity	-0.37	0.09	0.000 ***	-0.38	0.11	0.000 ***	-0.37	0.08	0.000 ***
6. Same identity	0.65	0.16	0.000 ***	0.63	0.17	0.000 ***	0.61	0.14	0.000 ***
7. Geodistance	0.02	0.05	0.716	0.02	0.06	0.766	0.02	0.05	0.708
8. Geoproximity	-0.03	0.05	0.503	-0.04	0.06	0.574	-0.04	0.05	0.472
Functional dimensions (exogenous)									
9. Role_max	-0.44	0.11	0.000 ***	-0.49	0.23	0.031 *	-0.42	0.10	0.000 ***
10. Same role_max	0.02	0.18	0.923	0.00	0.18	0.989	-0.02	0.16	0.878
11. Role_average	-0.68	0.20	0.001 ***	-0.59	0.27	0.028 *	-0.67	0.20	0.001 ***
12. Role_average similarity	-3.03	0.58	0.000 ***	-3.00	0.67	0.000 ***	-2.86	0.56	0.000 ***
Behavioral dimensions (exogenous)									
13. Role_sum	-0.01	0.03	0.716	0.00	0.07	0.984	-0.01	0.02	0.648
14. Role_sum similarity	0.01	9.06	0.999	-0.39	3.98	0.921	-0.34	8.68	0.969
15. Individual_sum	0.00	0.05	0.923	0.02	0.04	0.536	0.01	0.04	0.900
16. Individual_sum similarity	-4.35	9.62	0.651	-3.73	8.52	0.661	-4.42	9.32	0.635
Control variables									
17. Hotspots							0.78	0.32	0.017 *
Behavior Dynamics:									
0.3 Behavior (role_sum) rate period 1				704.36	94.60				
0.4 Behavior (role_sum) rate period 2				188.03	30.19				
18. Behavior (role_sum) linear shape				-0.06	0.02	0.004 **			
19. Behavior (role_sum) quadratic shape				0.00	0.00	0.003 **			
20. Behavior (role_sum) co_degree				0.00	0.00	1.000			
21. Behavior (role_sum) co_average alter				0.00	0.00	0.046 *			

The two-sided P-values were derived based on the normal distribution of the resultant test statistics (estimate devided by standard error). +p<.1, *p<.05, **p<.01, ***p<.001.

In Model 3, we have added a control variable to test if the effects identified in Models 1 are changed when we take into consideration the difference between hotspot projects and regular projects. The results show a statistically significant positive difference (beta = 0.78, P = 0.017), suggesting that organizations active in hotspots projects are more likely to form new collaborations over time than organizations that work in regular projects. The other effects remain similar.

All parameter estimates in the three models converged well below 0.1, indicating a good fit between the simulated ties and the observed ties. We also did sensitivity tests for the weighting of roles, but changing the weights did not influence the results. Overall goodness-of-fit (Figure 2) is with a p-value of 0.014, which is improved from 0.003 when only structural dimensions are included in the model. Most observations are nicely within the 95% regions of the simulated distributions, that indicates an acceptable fit of the models to the data.

5 CONCLUSIONS AND DISCUSSIONS

Stimulating and facilitating multi-actor collaborations for joint problem solving is considered to be one of the key challenges for modern organization studies. In practice, the emergence of new collaborative networks invariably entails a decision regarding who will participate and which partners to select. How organizations are connected can have lasting consequences for their performance. Yet, the mechanisms that may connect one actor to another remain insufficiently understood, owing to a restrictive focus on mechanisms of network endogeneity to the exclusion of exogenous mechanisms. In order to understand the mechanisms that influence the formation and evolution of collaborative networks, we have used a stochastic actor-based simulation model to study the evolution of a collaborative multi-actor program, combining endogenous and exogenous mechanisms of network formation.

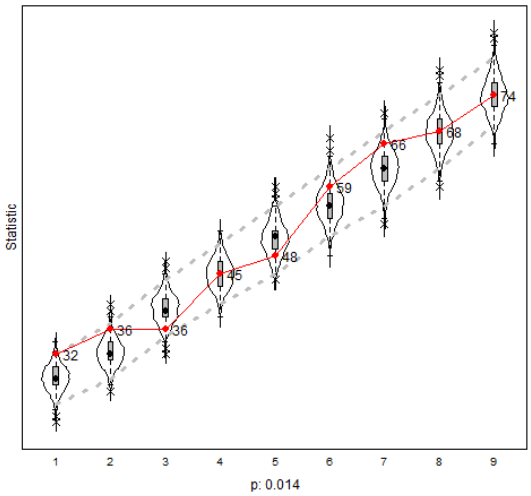


Figure 2. The goodness of fit of degree distribution.
The "violin plots" show, for each number of nodes with degree < x, the simulated values of these statistics as both a box plot and a kernel density estimate. The solid red line denotes the observed values. The dashed grey line represents a 95% probability band for the simulations.

The results of our analysis match the findings in previous literature with respect to endogenous network structural dimensions: transitivity and centrality play a key role in the evolution of the KvK network. The results also reveal the influence of exogenous mechanisms: actors tend to collaborate with other actors from the same type of organizations (componential) and patterns of collaboration are affected by the nature and differences in roles (functional), which may reflect task division within collaborative projects.

Our analysis reveals a gap between actors from different sectors and a gap between actors working on global problems and those working on local problems. The KvK research program was designed as platform to encourage and support the collaboration between actors from different sectors. The program aims to form a bridge between communities without necessarily closing the gap.

Our results also suggest that organizations active in hotspots projects are significantly more likely to form new ties than those active in theme projects. Hotspots projects focus on developing practical solutions for local and regional problems, while theme projects comprise teams of geographically dispersed scientists working to solve global challenges. The balance between global and local is reflected in the structure of the network.

Finally, our study has both theoretical and practical relevance. By addressing the mechanisms that inhibit or facilitate the development of collaborative networks, we provide theoretical insights in the position of organizations as strategic actors, attempting to effectively participate in organizational collaboration for knowledge creation. The practical value of our findings is that they may help identify and bridge gaps between actors from different societal organizations in a meaningful and purposeful way.

Our study is not without limitations, which also points the way for further research. First, we could only construct the presence or absence of ties (non-directed networks) from the available data. More information about who took the initiative to start a collaboration and other direction-related effects such as reciprocity would permit a more in-depth understanding and might also result in a better model fit. Second, the models were restricted to binary network data. Third, the project-based collaborations were affected by top-down (programme) interference for which we could not model. Finally, it would be interesting to investigate the emergent network at the individual level, which calls for a model with extended computational power.

REFERENCES

[1] L.M Camarinha-Matos and H. Afsarmanesh. Collaborative networks: a new scientific discipline. *Journal of Intelligent Manufacturing* 16: 439-452 (2005).

[2] P.R. Monge and N.S. Contractor. *Theories of Communication Networks*. Cambridge: Oxford University Press (2003).

[3] D. Stokols, K. L. Hall, et al. The science of team science: overview of the field and introduction to the supplement. *American Journal of Preventive Medicine* 35(2): S77-S89 (2008).

[4] K. Borner, N. Contractor, et al. A Multi-level systems perspective for the science of team science. *Science Translational Medicine* 2(49).

[5] G. Melin. Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy* 29(1): 31-40 (2000).

[6] M.J. Burger and V. Buskens. Social context and network formation: An experimental study. *Social Networks* 31: 63-75 (2009).

[7] H. Flap. Creation and returns of social capital: a new research program. In: *Creation and Returns of Social Capital: A New Research Program*, pp. 3-23. H. Flap, B. Völker (Eds.). Routledge, London (2004).

[8] L.M Camarinha-Matos and H. Afsarmanesh. A comprehensive modelling framework for collaborative networked organizations. *Journal of Intelligent Manufacturing* 18: 529-542 (2007).

[9] T.A.B. Snijders. The statistical evaluation of social network dynamics. *Sociological Methodology* 31: 361-395 (2001).

[10] T.A.B. Snijders. Models for longitudinal network data. In: *Models and Methods in Social Network Analysis*, pp. 215-247. P. Carrington, J. Scott, S. Wasserman (Eds.). Cambridge University Press, Cambridge (2005).

[11] T.A.B. Snijders, C.E.G. Steglich and M. Scheweinberger. Modelling the co-evolution of networks and behaviour. In: *Longitudinal Models in the Behavioural and Related Sciences,* pp.41-71. K. Montfort, H. Oud, A. Santorra (Eds.). Mahwah, NJ: Lawrence Erlbaum (2007).

[12] T.A.B. Snijders, G.G. van de Bunt and C.E.G. Steglich. Introduction to Stochastic Actor-based Models for Network Dynamics. *Social Networks* 32(1): 44-60 (2010).

[13] G.R. Carroll and M.T. Hannan. *The demography of corporations and industries*. Princeton, NJ: Princeton University Press (2000).

[14] D.G. McKendrick and G.R. Carroll. On the genesis of organizational forms: evidence fromthemarket for disk drive arrays. *Organization Science* 12: 661-683 (2001).

[15] P. Lazarsfeld and R.K. Merton. Friendship as a social process: a Substantive and methodological analysis. In: *Freedom and Control in Modern Society*, pp. 18-66. B. Morroe, T. Abel, C.H. Page (Eds.). New York: Van Nostrand (1954).

[16] J.S. Katz. Geographical proximity and scientific collaboration. *Scientometrics* 31(1): 31-43 (1994).

[17] M. Ruef, H.E. Aldrich and N.M. Carter. The structure of founding teams: homophily, strong ties, and isolation among U.S. entrepreneurs. *American Sociological Review* 68(2): 195-222 (2003).

[18] J. Skvoretz and T. Fararo. Status and participation in task groups: a dynamic network model. *American Journal of Sociology* 101: 1366-1414 (1996).

[19] M.H. Fisek, J. Berger, and R.Z. Norman. Participation in heterogeneous and homogeneous groups: a theoretical integration. *American Journal of Sociology* 97: 114-142 (1991).

[20] R. Berardo and J.T. Scholz. Self-organizing policy networks: risk, partner selection, and cooperation in Estuaries. *American Journal of Political Science* 54(3): 632-649 (2010).

[21] M. Checkley and C. Steglich. Partners in power: job mobility and dynamic deal-making. *European Management Review* 4: 161–171 (2007).

[22] G.G. van de Bunt and P. Groenewegen. Dynamics of collaboration in interorganizational networks: an application of actor-oriented statistical modelling. *Organizational Research Methods* 10: 463-482 (2007).

[23] J.J. Ebbers and N.M. Wijnberg. Disentangling the effects of reputation and network position on the evolution of alliance networks. *Strategic Organization* 8 (3): 255-275 (2010).

[24] P. Zappa. The network structure of knowledge sharing among physicians. *Quality and Quantity* 45: 1109-1126 (2011).

[25] J.J. Ebbers and N.M. Wijnberg. Disentangling the effects of reputation and network position on the evolution of alliance networks. *Strategic Organization* 8(3): 255-275 (2010).

[26] R.J. DeFillippi and M.B. Arthur. Paradox in project-based enterprise: the case of film making. *California Management Review* 40(2): 1-15 (1998).

[27] M. Bastian, S. Heymann and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media* (2009).

[28] J.A. Lospinoso and T.A.B. Snijders. Goodness of fit for Social Network Dynamics. Presentation given at the Sunbelt XXXI: St. Pete's Beach, Florida, USA (2011).

Epistemic Responsibility in Entangled Socio-Technical Systems

Judith Simon[1]

Abstract. In my talk I want to start exploring the requirements for a concept of epistemic responsibility that can account for the responsibilities of different (human and non-human) agents within entangled socio-technical epistemic systems. This includes the question as to whether non-human epistemic responsibility is possible in the first place or whether non-human agents can merely exhibit agency and accountability but no responsibility. To open up this topic, I will make use of insights from three different fields of research, namely: research on (distributed) moral responsibility in philosophy of computing, research on epistemic responsibility in (social) epistemology and research on distributed or entangled responsibility in feminist theory.

1 INTRODUCTION

Contemporary epistemic practices have to be conceived as socio-technical epistemic practices. That is, our ways of knowing, be it in research or in everyday-life are on the one hand highly social: much of what we know, we know through the spoken or written words of others; research consists not only in collaboration, but also in building upon previous knowledge, in communicating information, in communal quality assessment of scientific agents or content (e.g. peer review), etc. On the other hand, technology, particularly information and communication technologies mediate and shape these practices of knowing to profound extends. Social computing aligns these technical and social aspects. If we use social computing for epistemic purposes, we can speak of socio-technical epistemic systems par excellence: We check Wikipedia to find information about a city we plan to visit or some information about a historical incident, we rely on search engines to deliver relevant information on a specific topic, we use ratings of other agents explicitly to assess the quality of products before buying them or implicitly by accepting the ordering of search results or recommendations.

In knowing, we rely in numerous more or less transparent ways on other agents, human agents as much as non-human agents, infrastructures, technologies. However, this socio-technical entanglement in knowing is philosophically still only poorly understood. How do we trust to know – and how should we trust to know in socio-technical epistemic systems? What could epistemic vigilance mean – on the web and elsewhere? What are the epistemic responsibilities of different agents, e.g. of designers or users of search engines or recommender systems? How should concepts such as agency, accountability and responsibility in socio-technical epistemic systems and their epistemic counterparts be understood in the first place?

Different (sub-)disciplines have provided invaluable insights to crucial *aspects* of knowing within entangled socio-technical epistemic systems, even if none of them has yet offered any comprehensive account of it. Providing such a comprehensive account is beyond the scope of my talk. Hence, I want to focus on a more specific topic namely *epistemic responsibility.* More precisely, *the goal of my talk is to explore the requirements for a concept of epistemic responsibility that can account for the responsibilities of different (human and non-human) agents within entangled socio-technical epistemic systems.* This includes the question as to whether non-human epistemic responsibility is possible in the first place or whether non-human agents can merely exhibit agency and accountability but no responsibility.

To open up this topic, I will make use of insights from three different fields of research, which I will very briefly introduce in the following sections: research on *(distributed) moral responsibility in philosophy of computing,* research on *epistemic responsibility in (social) epistemology* and research on *distributed or entangled responsibility in feminist theory.*

2 RESPONSIBILITY & ICT: INSIGHTS FROM THE PHILOSOPHY OF COMPUTING

The difficulty to attribute responsibility, to locate accountability in ever more distributed and entangled socio-technical systems is one of the core experiences which seems to pervade many, if not all aspects of our contemporary environment. Think small - about the difficulties of finding and reaching the person to make responsible in case of a non-functioning internet connection? Think big – who's responsible for the financial crisis?

Computer technology and ICT in particular has deepened and aggravated these issues. Think of artificial agents, search engine algorithms, the personal data handling of social networking sites; think of drones, robots in military and health-care or unmanned vehicles: who is responsible, who is to blame if things go wrong: designers, users, the technologies or rather the distributed and entangled socio-technical systems in compounds?

There is a growing amount of research on moral and legal responsibility in computing (cf. [1]), specific foci being autonomous agents (e.g. [2]) and robotics [3]. With respect to accountability, Nissenbaum's paper [4] on accountability in a computerized society is surely an early seminal piece, in which different causes for difficulties in accountability attribution are worked out: the problem of many hands, the problem of bugs, using the computer as a scapegoat, and ownership without liability.

[1] Department of Philosophy, University of Vienna, Austria & Institute of Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Germany. Email: judith.simon@univie.ac.at.

Of particular importance for the goals of this paper are Floridi and Sander's early considerations on *the morality of artificial agents* and the concept of *distributed morality* [5]. According to them something qualifies as an agent if it shows *interactivity*, *autonomy* and *adaptability*, i.e. neither free will nor intentions are deemed necessary for agency. In the context of social computing, such a concept of "mind-less morality" [5: 349] allows addressing the agency of artificial entities (such as algorithms) as well as of collectives, which may form entities of their own (such as companies or organizations). Another merit of their approach lies in the disentanglement of moral agency and moral responsibility: a non-human entity can be held accountable if it qualifies as an agent, i.e. if it acts autonomously, interactively and adaptively. However, it cannot be held responsible, because responsibility requires intentionality. That is, while agency and accountability do not require intentionality, responsibility does. Therefore, it seems that non-human agents – at least in separation – cannot be held *responsible* even if they are *accountable* for certain actions. I will return to this topic at the end of this paper.

While these considerations on responsibility and accountability in socio-technical systems are highly developed, the specific problem of *epistemic* responsibility in ICT has not yet been in the focus of attention within philosophy of computing. Hence, to understand more about the specificities of epistemic responsibility, we should also turn to epistemology, and to social epistemology in particular.

3 EPISTEMIC RESPONSIBILITY: INSIGHTS FROM SOCIAL EPISTEMOLOGY

In social epistemology, debates concerning the epistemological status of testimony (e.g. [6], [7], [8],), have in the new millennium also led to explorations of the notions of epistemic trust (e.g. [9]), epistemic authority (e.g. [10]), epistemic injustice (especially [11]) – and now most recently also, epistemic responsibility.[3]

Due to this origin in the debates around the epistemology of testimony, the focus of attention in this discourse of epistemic responsibility is also mostly on epistemic interactions between human agents, i.e. on the responsibilities of speakers and hearers in testimonial exchanges. Yet, taking into account that processes of knowing take place in increasingly entangled systems consisting of human and non-human agents, systems in which content from multiple sources gets processed, accepted, rejected, modified in various ways by these different agents, the notion of epistemic responsibility needs to be modified and expanded to account for such epistemic processes. In particular, I think two issues need to be addressed in more detail than is currently the case in most analytic accounts of epistemic responsibility: a) the role of technology and b) the relationship between power and knowledge.[4] For both topics, feminist theoreticians in particular have provided highly valuable insights.

[3] Confer for instance the conference on "Social Epistemology and Epistemic Responsibility", which took place at Kings College in May 2012.
http://www.kcl.ac.uk/artshums/depts/philosophy/events/kclunc2012.aspx

[4] It would be inadequate to argue that the role of technology or the role of power have been entirely neglected in social epistemology. On the one hand, there have been attempts to account for ICT (e.g. some works by Alvin Goldman [14] and Don Fallis [15], the special issue of the journal EPISTEME (2009, volume 6, issue 1, on Wikipedia). Moreover, Fricker's [11] book on "Epistemic Injustice" has also stirred a lot of interest in the relationship between power and knowledge. However, these developments are rather recent and the classical assessment of testimonial processes remains focused on communication between humans often still conceived as an unconditioned and a-social subject *S, who knows that p.*

4 EPISTEMIC RESPONSIBILITY IN ENTANGLED SOCIO-TECHNICAL SYSTEMS: INSIGHTS FROM FEMINIST THEORY

Despite the fact that epistemic responsibility has only very recently attracted attention within analytic epistemology, the term itself has already been used in 1987 as the title of a book by Lorraine Code [12]. In this book, Code addresses the concepts of responsibility and accountability from a decidedly feminist perspective and argues that in understanding epistemic processes in general and epistemic responsibility and accountability in particular, we need to relate epistemology to ethics. Criticizing the unconditioned subject S who knows that p, "the abstract, interchangeable individual, whose monologues have been spoken from nowhere, in particular, to an audience of faceless and usually disembodied onlookers" [13:xiv], Code emphasizes social, i.e. cooperative and interactive aspects of knowing as well as the related "complicity in structures of power and privilege" [13:xiv],, "the linkages between power and knowledge, and between stereotyping and testimonial authority" [13:xv].

While Code's work highlights the relationship between knowledge and power, research by Karen Barad and Lucy Suchman adds technology to the equation and therefore appears particularly suited to explore the notion of epistemic responsibility within entangled and distributed socio-technical systems:

Barad's "agential realism" (AR) [16, 17], delivers an "[...] epistemological-ontological-ethical framework that provides an understanding of the role of human and nonhuman, material and discursive, and natural and cultural factors in scientific and other social-material practices" [17:26].

Barad's AR is theoretically based upon Niels Bohr's unmaking of the Cartesian dualism of object and subject, i.e. on the claim that within the process of physical measurement, the object and the observer, Barad's "agencies of observation", get constituted by and within the process itself and are not pre-defined entities. The results of measurements are thus neither fully constituted by any reality that is independent of its observation, nor by the methods or agents of observation alone. Rather, all of them, the observed, the observer and the practices, methods and instruments of observation are entangled in the process of what we call "reality". For Barad, reality itself is nothing pre-defined, but something that develops and changes through epistemic practices, through the interactions of objects and agents of observation in the process of observation and measurement. Reality in this sense is a verb and not a noun.

Yet, *inter*action is a problematic term in so far as it presupposes two separate entities to interact. Thus, to avoid this presupposed dualism, she introduces the neologism of "intra-

action", to denote the processes taking place within the object-observer-compound, the entanglement of object and observer in the process of observation. This terminological innovation is meant to discursively challenge the prevalent dualisms of subject-object, nature-culture, human-technology, and aims at opening up alternative, non-dichotomous understandings of technoscientific practices.

A crucial concern of Barad is the revaluation of *matter*. Opposing the excessive focus on *discourse* in other feminist theories (e.g. Judith Butler's), Barad emphasizes the relevance of matter and the materiality of our worlds. Taking matter serious and describing it as active, means to allow for non-human or hybrid forms of agency, a step that has been taken already with the principle of general symmetry in Actor-Network-Theory. But then here is the problem: If we attribute agency to non-human entities, can and should they be held responsible and accountable? Plus, isn't that an invitation, a carte blanche to shirk responsibility by humans? Do we let ourselves off the hook to easily and throw away any hopes for responsible and accountable actions?

It appears that Barad's view on non-human agency and her stance towards the ontological asymmetry between humans and non-humans has changed from earlier articulations [16] to later ones [17]. In 1996, she still underscores the human role in representing, by stating that „[n]ature has agency, but it does not speak itself to the patient, unobstrusive observer listening for its cries – there is an important asymmetry with respect to agency: we do the representing and yet nature is not a passive blank slate awaiting our inscriptions, and to privilege the material or discursive is to forget the inseparability that characterizes phenomena" [16:181] .

However, it seems that this special treatment of humans and especially the notion of *representing* does not well match her posthumanist performativity, as depicted some years later [18]. Finally, in "Meeting the Universe Halfway" Barad offers a more nuanced dissolution of the distinction between human and non-human agency. By stating that "[a]gency is a matter of intra-acting; it is an enactment, not something that someone or something has" [17:261], Barad moves the locus of agency from singular entities to entangled material-discursive apparatuses. But even if agency is not tied to individual entities, it is bound with responsibility and accountability, as Barad makes very explicit in the following quote: "Learning how to intra-act responsibly within and as part of the world means understanding that we are not the only active beings— though this is never justification for deflecting that responsibility onto other entities. The acknowledgment of "nonhuman agency" does not lessen human accountability; on the contrary, it means that accountability requires that much more attentiveness to existing power asymmetries [17:218f].

Thus, the possibility to understand agency not essentialist as a (human) characteristic, but as something which is rather *attributed*[5] to certain phenomena within entangled networks *could* be regarded as an invitation to shirk of responsibility. But this is clearly not the case for Barad. When developing her posthumanist ethics, Barad concludes that even if we are not the only ones who are or can be held responsible, our responsibility even greater than it would be if it were ours alone. She states "We (but not only "we humans") are always already responsible to the others with whom or which we are entangled, not through conscious intent but through the various ontological entanglements that materiality entails. What is on the other side of the agential cut is not separate from us—agential separability is not individuation. Ethics is therefore not about right response to a radically exterio/ized (sic!) other, but about responsibility and accountability for the lively relationalities of becoming of which we are a part." [17:393].

This focus on responsibility and accountability relates back to Barad's initial framing of agential realism as an "epistemological-ontological-ethical framework", a term by which she stresses the "[...] fundamental inseparability of epistemological, ontological, and ethical considerations" [17:26]. Barad insists that we are responsible for what we know, and – as a consequence of her onto-epistemology for what is [18:829]. Accountability and responsibility must be thought of in terms of what matters and what is excluded from mattering, what is known and what is not, what is and what is not.

This acknowledgement that knowledge always implies responsibility, not only renders issues of ethics and politics of such knowledge- and reality-creating processes indispensable. It also relates directly back to Barad's emphasis on performativity. Epistemic practices are productive and different practices produce different phenomena. If our practices of knowing do not merely represent what is there, but shape and create what is and what will be there, talking about the extent to which knowledge is power or entails responsibility gets a whole different flavor.

Lucy Suchman shares many concerns of Barad and her insights promise to be of particular importance for social computing due to Suchman's background in Human-Computer Interaction. Acknowledging the relational and entangled nature of the sociomaterial, Suchman claims that agency cannot be localized in individual entities, but rather is distributed within socio-material assemblages. Resonating with Barad, she notes "[...] agencies – and associated accountabilities – reside neither in us nor in our artifacts but in our intra-actions" [19:285].

The question, however, remains how exactly to be responsible, how to hold or to be held accountable if agency is distributed. How can we maintain responsibility and accountability in such a networked, dynamic and relational matrix? Although I think that Suchman goes into the right direction, she remains quite vague about this in her concluding remarks of *Human-Machine-Reconfigurations* by stating that „responsibility on that view is met neither through control nor abdication but in ongoing practical, critical, and generative acts of engagement. The point in the end is not to assign agency either to persons or to things but to identify the materialization of subjects, objects, and the relations between them as an effect, more and less durable and contestable, of ongoing sociomaterial practices" [19:285].

5 DISENTANGLING (EPISTEMIMC) AGENCY, ACCOUNTABILITY AND RESPONSIBILITY

To understand the epistemic responsibilities of knowers in our contemporary world, I think all insights outlined above need to be accounted for. Yet it still has to be discussed in detail whether, how and to what extent they can be aligned. As knowers we move and act within highly entangled socio-

[5] Cf. Wallace (1994) on the attribution of responsibility.

technical epistemic systems. In our attempts to know, we permanently need to decide when and whom to trust and when to withhold trust, when to remain vigilant. Loci of trust in these entangled and highly complex environments are not only other humans, but also of technologies, companies, or organizations – and they usually cannot be conceived in separation but only as socio-technical compounds.

However, the fact that both human and non-human entities can qualify as agents should not convey the impression that we have entered a state of harmony and equality: there are enormous differences in power between different agents. To use Barad's terminology, some agents *matter* much more than others. And those that matter most do not necessarily have to be the human agents.

Socio-technical epistemic systems in general and social computing applications in particular need to be understood as highly entangled but also highly differentiated systems consisting of human, non-human and compound or collective entities with very different amounts of power. To understand this, search engines are a useful example. In highly simplified terms, search engines can be conceived as code written, run and used by human and non-human agents embedded in socio-technical infrastructures as well as in organizational, economic, societal and political environments. While there are potentially many ways to enter the World Wide Web, search engines have emerged as major points of entrance and specific search engines nowadays function as "obligatory passage points" [20], exerting enormous amount of not only economic, but also epistemic power.

What do these considerations and insights imply for the development of a useful concept of epistemic responsibility? First of all, it should be noted that responsibility is something that can be assumed oneself as well as something that can be attributed to someone or something else. This dual nature of responsibility has to be kept in mind if we want to understand what it means to be epistemically responsible, because we can ask two questions: 1) Can epistemic responsibility be *assumed* only by human agents or also by other agents? 2) Can epistemic responsibility be *attributed* to only human or also non-human agents? Or are these two questions already misleading because they imply or at least allow for individualized forms of responsibility, which appear at odds with Barad's view. Irrespective of how we respond to this, these questions would be starting points for inquiry at maximum, because the next steps would then consist in finding criteria of *how exactly* responsibility can be assumed or attributed and further how it *should* be assumed or attributed.

To my mind, a first step should consist in disentangling the notions of agency, accountability and responsibility more carefully. While both Barad and Suchman in the previous quotes seem to use the terms synonymously, it seems fruitful to keep up a distinction – in particular, to understand both notions and their epistemic counterparts in entangled socio-technical systems. For this distinction between responsibility and accountability, insights from computer ethics can be of some use – even if different premises may lead to some initial contradictions, which would need to be resolved by further research. According to Floridi and Sanders [5], agency requires only interactivity, autonomy and adaptivity, but no intentionality is needed. Accountability is bound to agency only and hence also does not require intentionality of agents. However, responsibility differs from accountability exactly by requiring intentionality. Hence, if we agree with Floridi and Sanders [5] that responsibility as opposed to agency and accountability requires intentionality, then it makes no sense to talk about responsibility with respect to technical artifacts. A car cannot be made responsible for a crash, it is the driver who is to blame - for negligence or ill-will - or maybe the manufacturer, if a technical flaw caused the crash. Even if we think of unmanned vehicles and the car drove autonomously, interactively and adaptively and then caused a crash, this car may be *accountable* for a crash, but it could not be made *responsible*. Please note that it is only the technical artifact *in isolation*, which cannot be made responsible. For socio-technical compounds, the possibility of attributing responsibility would still be given, hence this perspective may in the end well be compatible with Barad's agential realism [17].

If we want to distinguish responsibility and accountability than sticking to intentionality as the demarcation line appears still plausible and fruitful. Moreover, I think the same distinctions between agency, accountability and responsibility also hold for their epistemic counterparts: algorithms, software applications or interfaces may have epistemic agency and then could be made epistemically accountable, but it is unclear how they – in isolation – could be considered *responsible* in a strong sense of the word which differentiates between accountability and responsibility.

For responsibility to be attributed some human (either individually or as part of a collective) seems to have to be part of the socio-technical compound. Both Barad and Suchman have reminded us that analytic *cuts* are never innocent, that the distinctions we make and boundaries we draw in research have consequences and should therefore be done carefully. This does not imply, however, that cuts can be avoided, that they should not or cannot be done for epistemic purposes. Hence, I consider it adequate to also take a look cut-out or individualized agents. Even if we acknowledge the thorough entanglement of agents, we may need to zoom in and cut out parts of this entanglement not only to understand more about this part, but also about its surroundings. And as we have seen, even those cut-out parts, already pose enormous conceptual and pragmatic difficulties. Nonetheless, the task remains to tackle the responsibility of socio-technical compounds. If we decide to keep intentionality as the demarcation line between responsibility and accountability, insights from the field of social ontology, especially debates on *shared intentionality* and *group agency* may prove useful [21, 22, 23].

5 OUTLOOK

In my talk, I hope to further expand and deepen these initial considerations concerning the problems related to epistemic responsibility within distributed socio-technical systems and to explore how these insights can be made fruitful for social computing. While it is clear, that providing full-blown models or definitive answers of how to conceive epistemic responsibility in socio-technical epistemic systems is beyond the scope of such a short paper, I hope to open up a new field of inquiry, to have asked questions that will lead to new insights.

REFERENCES

[1] Coleman, K. G. (2004). "Computing and Moral Responsibility." Stanford Encyclopedia of Philosophy. from http://plato.stanford.edu/entries/computing-responsibility/.

[2] Coeckelbergh, M. (2009). "Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents." AI & Society 24(2): 181-189.

[3] Pagallo, U. (2010). "Robotrust and legal responsibility." Knowledge, Technology & Policy 23(3-4): 367-379.

[4] Nissenbaum, H. (1997). Accountability in a Computerized Society. Human Values and the Design of Computer Technology. B. Friedman. Cambridge, Cambridge University Press: 41-64.

[5] Floridi, L. and Sanders J.W. (2004). "On the morality of artificial agents." Minds and Machine 14: 349-379.

[6] Coady, C. A. J. (1992). Testimony. A Philosophical Study. Oxford, Claredon Press.

[7] Fricker, E. and D. E. Cooper (1987). "The Epistemology of Testimony." Proceedings of the Aristotelian Society 61(Supplementary Volumes): 57-106.

[8] Adler, J. (1994). "Testimony, Trust, Knowing." The Journal of Philosophy 91(5): 264-275.

[9] Origgi, G. (2004). "Is trust an epistemological notion?" Episteme 1(1): 1-12.

[10] Origgi, G. (2008). "Trust, authority and epistemic responsibility." Theoria 61: 35-44.

[11] Fricker, M. (2007). Epistemic Injustice. Power and the Ethics of Knowing. Oxford, Oxford University Press.

[12] Code, L. (1987). Epistemic Responsibility. Hanover, New England, University Press of New England.

[13] Code, L. (1995). Rethorical Spaces: Essays on Gendered Locations, Routledge.

[14] Goldman, A. I. (2008). The Social Epistemology of Blogging. Information Technology and Moral Philosophy. J. v. d. Hoven and J. Weckert. New York, Cambridge University Press: 11-122.

[15] Fallis, D. (2006). "Social Epistemology and Information Science." Annual Review of Information Science and Technology 40: 475-519.

[16] Barad, K. (1996). Meeting the Universe Halfway. Realism and Social Constructivism without Contradiction. Feminism, Science, and the Philosophy of Science. L. H. Nelson and J. Nelson. Dordrecht, Holland, Kluwer: 161-194.

[17] Barad, K. (2007). Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning. Durham, Duke University Press.

[18] Barad, K. (2003). "Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter." Signs: Journal of Women in Culture and Society 28(3): 801-831.

[19] Suchman, L. A. (2007/2009). Human-Machine Reconfigurations. Plans and Situated Actions. Cambridge, Cambridge University Press.

[20] Callon, M. (1986) "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay", in: J. Law (ed.) Power, Action and Belief: A New Sociology of Knowledge, London: Routledge & Kegan Paul: 196-233.

[21] List, C. and Pettit, P. (2011) Group Agency. The Possibility, Design, and Status of Corporate Agents. New York: Oxford University Press.

[22] Gilbert, M. (2000) Sociality and Responsibility. Blue Ridge Summit: Rowman and Littlefield.

[23] Tollefsen, D. 2003a. "Collective Epistemic Agency." Southwest Philosophy Review 20(1): 55-66.

Trust in Social Machines: The Challenges

Kieron O'Hara[1]

Abstract. The World Wide Web has ushered in a new generation of applications constructively linking people and computers to create what have been called 'social machines.' The 'components' of these machines are people and technologies. It has long been recognised that for people to participate in social machines, they have to trust the processes. However, the notions of trust often used tend to be imported from agent-based computing, and may be too formal, objective and selective to describe human trust accurately. This paper applies a theory of human trust to social machines research, and sets out some of the challenges to system designers.

1 INTRODUCTION

Computers have always been sociotechnical systems, embedded in organisations, or serving the purposes of users for work or leisure. However, thanks to the spread of interactive read/write technologies (e.g. wikis, photo-sharing, blogging) and devices and sensors embedded in both physical and digital worlds (e.g. GPS-enabled hand-held devices), people and machines have become increasingly integrated. Terms such as 'augmented reality' and 'mediated reality' are in common use, and the embedding of computation into society via personal devices has led to discussion of *social machines* and *social computation*, an abstract conception in which people and machines interact for problem-solving. The 'components' of the machine may be people or computers; the 'routines' or 'procedures' could be carried out by humans, computers or both together.

Social machines are rapidly becoming a focus of computing research [1]. 'Programming the global computer' is one of the British Computing Society's grand challenges for computing, while peer-to-peer technologies have opened up the possibility of flexibly linking people and computers, as explored in projects such as OpenKnowledge (http://www.openk.org/) and the Social Computer community (http://www.socialcomputer.eu/).

Trust has always been recognised as an important factor in the function of such human/computer hybrids. However, the notions of trust used have often been relatively formal, imported from agent-based research. In this paper, I will examine the question of whether, and how, social computing can take into account wider and less well-ordered notions of psychologically realistic trust. I also note here two important limitations of scope of this paper. First, I focus here on issues of trust relevant to system designers fostering trust in their systems by users; of course there are many other stakeholders and many other trust relations typically involved (to take an obvious example, system designers have to trust users as well as being trusted by them). Secondly, I focus here on the challenges; solutions are already being created for these issues, but the point I want to emphasise in this paper is that we have to be clear about exactly how social machines rely on trust to function, and where a breakdown will lead to dysfunction. Without a precise model, it will be harder to diagnose problems.

[1] Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom, kmo@ecs.soton.ac.uk.

2 SOCIAL MACHINES

In this section, I will flesh out the idea of a social machine or social computer. After a preliminary discussion, I shall briefly describe a couple of examples. A third subsection will examine the notion of programming social machines, before the section is completed with a brief sketch of the important role trust plays.

2.1 What is a social machine?

The idea of a social machine was implicit in early conceptions of the World Wide Web. As Berners-Lee put it in 1999:

Real life is and must be full of all kinds of social constraint – the very processes from which society arises. Computers can help if we use them to create abstract social machines *on the Web: processes in which people do the creative work and the machine does the administration.* ([2], pp.172, Berners-Lee's emphasis)

We see plenty of social machines around today. Many are embedded in social networks such as Facebook, in which human interactions from organising a birthday party to interacting with one's Member of Parliament are underpinned by the engineered environment. Another type of example is a multiplayer online game, where a persistent online environment facilitates interactions concerning virtual resources between real people. A third type is an online poker game, where the resources being played for are real-world, but where the players may be human or bots, and where the environment in which the game takes place is engineered around a relatively simple computational model. In such systems, (some of) the social constraints that Berners-Lee talks about, which are currently norm-driven, are converted to (or in his terms administered by) the architecture of the programmed environment.

These social machines are straightforward (*qua* interaction models), but as the technology is theorised more deeply it is inevitable that more complex systems will be developed. A generalised definition of a social computation is provided by Robertson and Giunchiglia:

A computation for which an executable specification exists but the successful implementation of this specification depends upon computer mediated social interaction between the human actors in its implementation. [3]

In such an environment, self-organisation (partial or full) becomes viable and scalable, while physical objects, agents, contracts, agreements, incentives and other objects can be referred to using Web resources (Uniform Resource Identifiers –

URIs). 'Programming' the social computer (rather than simply supporting and directing interactions on an engineered environment) and integrating larger numbers of people and machines will become increasingly feasible.

2.2 Examples

As a small example of a social machine, consider reCAPTCHA [4]. A CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart), invented by Louis Von Ahn, is the distorted sequence of letters that someone has to type in a box to identify him- or herself as a human (e.g. to buy a ticket online, or to comment on a blog). This is a task that computers cannot do, and so the system stops bots buying thousands of tickets for a concert or sporting event for later resale, or for a spambot to leave spam messages as comments to blogs.

Von Ahn extended the idea of the CAPTCHA to create the reCAPTCHA, which uses the same principle to solve another problem. Google (which acquired reCAPTCHA in 2009) wishes to scan and publish out-of-copyright books. However, Optical Character Recognition is too fallible to automate the process (in books over 100 years old, OCR fails for about 30% of words). The quantity of books to be scanned rules out human labour as a general solution to the problem.

Von Ahn noticed that his original CAPTCHA device was being used over 200m times a day, about half a million person-hours of effort. reCAPTCHA was designed to put these person-hours to more productive use. It presents the user who wishes to identify him- or herself as a human with two words, not one. The first is a normal CAPTCHA, and the second is a word from an old book that OCR had failed to identify. If the person succeeds with the first CAPTCHA, then he or she is known to be a human. As humans are reliable at word recognition, Google can therefore take the response to the second word as a plausible suggestion of what it is. Presenting the same word to multiple users allows a consensus to emerge.

The person is not necessarily aware that he or she is helping Google in its scanning task. The incentive for his or her involvement is the need for identification (to buy tickets, or comment on a blog, etc). The time taken for a reCAPTCHA is not significantly longer than a CAPTCHA. The 'machine' thereby created, of millions of people interacting via the reCAPTCHA facility, is currently identifying about 100m words per day (about 2m books equivalent per year). reCAPTCHA is offered as a free Web service to hundreds of thousands of websites (including Facebook, Twitter and Ticketmaster) which need spam protection; the service can be offered without a fee because of the translation service it also provides to Google [4].

As another example, Robertson and Giunchiglia [3] use the DARPA balloon challenge of 2009, in which all human 'components' of the machine are fully aware of their own role. In the DARPA challenge, the aim was to find ten weather balloons placed randomly around the US (in nine different states from California to Delaware). The rules of the challenge were intended to support the growth of a network of people taking part in the search, enabling a crowdsourced solution. The means of doing this in the winning solution (from Sandy Pentland at the Massachusetts Institute of Technology) was to set out financial incentives – everyone who discovered a balloon got a certain quantity of money, while for everyone who received a reward, the person who introduced them to the network received half that reward. Hence people were incentivised both to look for the balloons and to add more people to the network. Pentland's team began with 4 people, and using social media had recruited over 5,000 at the point of completion, which took under ten hours.

reCAPTCHA and the DARPA challenge were intended to solve a particular exogenous problem, but social machines can be designed to solve the problems of the people who constitute them. In such cases, the incentive of the participants is that the machine's smooth functioning is in their own interests. One could imagine, for instance, a set of computer-mediated interactions enabling a community to provide a social response to problems of crime (such as BlueServo, which crowdsources the policing of the Texas-Mexico border), or enabling those suffering from a particular health care problem to pool resources and to offer support and advice to fellow sufferers (such as curetogether.com). It will be obvious from these examples that such efforts will not always be uncontroversial.

Note finally that in many cases the ability to compute and to gather and process information at large scale is vital. This adds an extra layer of complication to the social machine vision.

2.3 Programming the social machine

Giunchiglia and Robertson define a social machine or computer as follows [3]:

A computer system that allows people to initiate social computations (via executable specifications) and adopt appropriate roles in social computations initiated by others, ensuring while doing so that social properties of viable computations are preserved. A general purpose social computer provides a domain-independent infrastructure for this purpose.

This implies three processes that need to take place in order for the social machine to run. First, specifications must be *initiated*, so that where necessary groups of people are able and willing to carry out parts of the computation. It may be that part of the 'programming' of the social machine will involve observation of and induction from existing social processes, to be adapted and reused in the new context of the social machine.

Second, people and groups must *adopt appropriate roles* in the machine, having been incentivised to join social computations. The *discovery* of these roles is an important issue.

Third, the groups relevant to the computation must be *reinforced*; as Robertson and Giunchiglia put it, "this relies on the computation being executed in a way that spreads the computation and knits together the social group via further social properties of the computation." In other words, the social computation must preserve the social structures necessary for its operation. In the example of the DARPA challenge, the clause that rewards anyone who has introduced a reward-winner gives incentives to people to add friends to an ever-growing network.

Robertson and Giunchiglia also define a *social property*, analogous to an *invariant* in conventional programming with real-world physical consequences: "a requirement associated with the specification of social computation that must be maintained, and perhaps communicated, during the execution of the specification in order for the computation to establish the social group needed to run it."

So if we return to the example of reCAPTCHA, its initiation involves publicising the Web service to sites needing spam protection, people adopt the appropriate role when they decide to solve a reCAPTCHA to get access to a service, and relevant

groups are reinforced by the success of the service in suppressing spam on sites to which people want access. The key social property to be preserved is that spam is suppressed; if spammers found an effective way around the reCAPTCHA, then fewer sites would support the Web service, and therefore fewer people would be playing the role of word recognisers.

2.4 The relevance of trust

Trust is essential to the smooth running of a social machine. Two precondition for social machines to motivate people to adopt appropriate roles is that they trust that promised incentives will appear, and that they trust that the machine will not do anything (in the world) that conflicts with their values. In the case of reCAPTCHA, people must trust that they will obtain access to their desired sites. In the case of the DARPA challenge, the participants must have trusted that the money would be paid out.

Trust is also central to the reinforcement of groups, as cooperation towards a goal demands trust in others' contributions; would Wikipedia authors bother to contribute if their work was routinely trashed without argued rationales? If an effect of a computation was to fragment the coalitions developed to carry it out by undermining trust between members, then it could not ultimately succeed. It is fair to say that for many social computations, trust (both between individuals in different roles, and of the machine by its component individuals) is likely to be a social property essential to the social machine's function.

Trust is of course most important when people take risks or place themselves in a vulnerable position with respect to a social machine. With reCAPTCHA this is barely an issue, but in a machine that, for example, enabled people to manage health care problems, users might need to pool information which could include sensitive health- or lifestyle-related data. That brings in complex rights-based issues such as privacy, and legal issues such as data protection.

In the next section, I shall briefly set out some of the most important properties of trust, as background to a discussion of issues that arise with respect to trust in social machines.

3. TRUST

The discussion of trust will be in four parts, beginning with an analysis of trustworthiness, upon which will be built an analysis of trust. Finally I shall discuss issues surrounding the connection of the two. These analyses are developed in more detail in a working paper [5].

3.1 Trustworthiness

Trustworthiness is prior to trust, which is an attitude toward the trustworthiness of others. Indeed, as Hardin has argued ([6], [7]), many commentators supposedly discussing trust are actually discussing trustworthiness. What, then, is this prior concept?

A trustworthy person is someone who does what she says she will do, all things being equal. This characterisation conceals quite a lot of structure. First of all, trustworthiness is a *property* of an *agent*. A *claim* must be made about her future actions. After all, it is absurd to accuse Barack Obama of being an untrustworthy brain surgeon, because he has never claimed to have brain surgery skills. The claim will also narrow the scope of trustworthiness; put another way, trustworthiness is context-dependent. The 'all things being equal' clause means that a trustworthy person need not succeed in carrying out the claimed behaviour, but if she does not, there must be an explanation for her failure which will absolve her of responsibility.

We can therefore define trustworthiness as a four-place relation, as follows:

(1) Y is trustworthy $=_{df}$ Tw<Y,Z,R,C>

where Y and Z are agents, R is a representation of the claim and C is a (task) context in which it applies.

In (1), Y is the agent who, if (1) is true, is trustworthy. R is the content of the claim made about her intentions, capacities and motivations for future behaviour. When (1) is true, Y's behaviour will be constrained by R. R may be explicitly written down, or may be implicit and understood; it may be open-ended and deliberately left unspecific to degrade gracefully. C is the set of contexts in which R is intended to apply (for instance, Y may claim to be a trustworthy car mechanic, but only within office hours, and only on certain makes of car).

This leaves Z, who is the agent responsible for generating and disseminating the claim R. In many, perhaps most, circumstances, Y = Z. However, this need not be the case. A trustworthy customer service employee (Y) respects a role description generated by her company (Z). A trustworthy piece of software (Y) performs according to a specification written by a designer (Z). It is essential that Z is *authorised* to make the claim about Y. Without authority, Z's claim has no bearing on Y's trustworthiness.

3.2 Trust

Trust is an *attitude* toward the trustworthiness of another. X trusts Y iff he believes that she is trustworthy (or, better, holds of the proposition 'Y is trustworthy' that it is true).

This characterisation of trust has a straightforward surface appearance. It is still a complex idea, however. Not only does trustworthiness import context-dependency, but trust forces us to confront a subjective element. There are six parameters of consequence in the trust relation, as follows:

(2) X trusts Y $=_{df}$ Tr<X,Y,Z,I(R,c),Deg,Warr>

with Y, Z and R as before, and X an agent.

In (2), the first three parameters are the relevant agents. X is the trustor and Y the trustee. Z, as before, is the agent who makes the claim R about Y's intentions, capacities and motivations. And again, as before, it could be that Z = Y (or, for that matter, X = Y, X = Z or X = Y = Z, although the possibility of these identities will not be defended here [5]).

Z makes a claim that Y's behaviour, all things being equal, will conform to R in contexts C. X's trust, if well-placed, should accept that claim. However, it need not, because X is only boundedly rational and communications between Z and X are not guaranteed to succeed. Furthermore, R might be implicit or unspecific. Hence X has to *interpret* R's meaning in the contexts in which he is interested. I have written this as a function I(R,c), to be read as X's interpretation of the force of R in the set of contexts that interest X, which I term c.

This brings trust's subjective aspect to the fore. For X's trust, it is *X's interpretation* that is the final arbiter, whether or not it is accurate. As trust is an attitude held by X about Y, it is X who supplies the underlying assumptions of the judgment. This has three specific consequences. First, for Y to maintain X's trust,

she must behave in accordance with I(R,c) *even if that differs from her own interpretation of R in c*. Second, for X to trust Y, it need not be the case that Z *has* authority to make claim R about Y. It is necessary only that X *believes that* Z has that authority. Third, I(R,c) only has any force with respect to Y if $c \subseteq C$, otherwise it will fall out of the scope of R. Yet for X's trust, it is necessary only that X *believes that* $c \subseteq C$. If any of X's beliefs is false – i.e. if the force of R in c is not I(R,c), or if Z does not have the authority to make claim R about Y, or if $c \nsubseteq C$ – X's trust or mistrust will be misplaced as based on a misunderstanding.

In short, in definition (2) above, X *believes that* (i) Z can authoritatively make claim R about Y, (ii) I(R,c) is the interpretation of R within a set of contexts c, and (iii) $c \subseteq C$.

This leaves two more parameters. Deg is a measure of X's confidence in his attitude toward Y's trustworthiness. The metric for Deg depends on the system under discussion. For psychological realism, it may be that Deg would be a fairly coarse-grained Likert-type psychometric scale of five or seven points. But it would be legitimate to produce more complex models that modelled Deg on, say, the real line between 0 and 1.

Whatever metric chosen must facilitate the expression of two types of trust judgment. First of all, X may have to choose whether he trusts Y_1 more than Y_2 to decide with whom to place his trust. Secondly, the level of risk that X takes on with respect to an interaction with Y will depend on his degree of trust; if he trusts her a lot, he will, all things being equal, be prepared to risk a lot, and if he trusts her only a little, his appetite for risk will be diminished.

Warr is the warrant for X's trust in Y. This could take any form – it doesn't have to be rational, and could even be that X has been dosed with oxytocin which increases the propensity to trust [8]. Unlike a warrant in Toulmin's system [9], the warrant explains the judgment, but is not intended for the persuasion of others. Nevertheless, usually there is a sensible rationale behind a trust judgment which is important for assessing it, and also for assessing how robust it is likely to be. Typical relatively reliable trust warrants include the reputation of Y, the past history of X's encounters with Y, the availability of sanctions for X, the possibility of a binding reciprocal agreement between X and Y, the credible commitments made by Y and the credentials that Y brings to the transaction.

As Wierzbicki argues ([10], pp.26-27), trust that does not have a rational component will be hard to model. That does not mean that trust cannot be irrational, but it makes it harder to embed psychologically-realistic trusting mechanisms into software, or to design sociotechnical systems (or social machines) which incorporate potentially irrational human trust judgments without restriction.

3.3 The problem of trust

The problem of trust is not to increase trust, but rather to ensure that X trusts Y when and only when Y is trustworthy. This is difficult as the incentives are not optimally aligned. If X risks assets in an interaction with Y, then he *benefits* from her *trustworthiness*, but unfortunately he only *controls* his *trust*. Conversely, Y benefits from X's trust, but only controls her trustworthiness. The result is a dilemma where the benefits of cooperation could be high, but losses to a trusting (trustworthy) party would accrue if their partner is untrustworthy (distrusting).

From this two things follow. First, trust cannot be an entirely rational attitude; as Hollis has argued, trustworthiness does not survive rigorous game-theoretic analysis (a fact available to rational would-be trustors) [11]. Second, X should use the analysis of (2) to determine where trust judgments can break down. Many failures of trust are down to differences in interpreting what Y is committed to.

A typical strategy for a trustworthy Y is to send *signals* of trustworthiness to X, which ideally will accurately represent her trustworthiness (would not be forthcoming if she were *not* trustworthy) and which will be included in X's warrant to trust Y [12]. These signals can be conscious or unconscious, and more or less strongly connected with the task that Y is offering to carry out, preferably as an unavoidable by-product. The flip side of any such signalling system, however, is that if it is made explicit, it can potentially be counterfeited by an untrustworthy person. Types of signal already mentioned include Y's reputation, history and credible commitments.

A second strategy involves structuring the encounter with some kind of *institution* (in the broad sense of a mechanism for producing order by structuring behaviour) which can reduce the likelihood of a deception being in Y's interest. Such an institution might supply objective credentials for Y, or might make plausible and effective sanctions available for X to apply if Y defects. Or X and Y might set up their own 'mini-institution' by entering into a reciprocal agreement. In each case, an institution promotes X's trust in Y only if X trusts the institution to deliver the structures it promises.

4 TRUST IN SOCIAL MACHINES: CURRENT APPROACHES

As noted earlier, trust is a vital element for social machines to function. However, this is a complex issue: in the open peer-to-peer architectures that will be required to support social machines, traditional knowledge engineering safeguards (such as centralisation of key functions, shared culture and ontologies, constraints and access control) are not practicable. In this section, I will expand on the theme of trust, using the theoretical apparatus assembled in Section 3.

Importing human interaction into the programming environment envisaged by Robertson and Giunchiglia presents a major challenge. Hendler and Berners-Lee see artificial intelligence as the key to enable people and machines to represent and reason over social attitudes including trust and trustworthiness, as well as related issues such as reliance and expectations; linked data and the Semantic Web will be important tools in such a world, by providing designers with access to a level of abstraction in which resources can be referred to directly and independently of the documents in which they are described [13]. Machines which require users to contribute information (such as those mentioned earlier to coordinate community responses to crime or healthcare issues) will also need to reason about privacy and data protection.

The human world is messy and full of compromise; computations in social machines must be able to cope with the consequences of this, such as inconsistency. Furthermore, given the sensitivity of personal data, social machines will also need to be able to function in hostile environments where some actors are malicious.

Although this is a lively area for research, there are few robust and scalable structures in place to represent these qualities. Hendler and Berners-Lee point out the importance of being able to treat these social phenomena as first-class objects capable of being reasoned over. The Semantic Web provides a blueprint for this, allowing the use of URIs to name objects of any kind [13].

In open environments, trust needs to be fostered from a number of sources. The most common view is to describe the relations between peers in a peer-to-peer architecture in terms of permissions and obligations governed by policies [14]. Theorem provers can determine whether peers have conformed to policies [15] and systems have been developed to explore the question of how to specify and verify strategies to determine whether and when to interact, and with whom [16].

5 DISCUSSION: THE HUMAN ELEMENT

One issue is that these approaches tend to assume that human trust behaviour is relatively well-behaved and if not rational at least fairly tidy and explicable. Yet as argued in section 3, it need not necessarily be so; as Kahneman has recently pointed out, rational processing coexists with fast, intuitive and emotional thinking [17]. Furthermore, the subjective element of trust is deep-seated. Hence policies may work very well to describe interactions in distributed systems unless elements are likely to behave idiosyncratically. Reasoning is only one approach to making a trust judgment, and may well involve a complexity that is inappropriate. Human judgments about trustworthiness of complex and distributed systems will not always align with the methods, ontologies and terms in which questions are framed by system designers. The key factors for consideration, as argued in section 3.2, include X's view of Z, X's interpretation of R, and the warrants that X accepts.

5.1 Displacing trust

Most approaches to trust in multi-agent systems assume that information relevant to agents' reputation, or data provenance, or data security will suffice to align trust and trustworthiness. Certainly transparency and availability of information about these is a bonus, and can do no harm. But will they be sufficient?

Trust is not always grounded; X's trust of Y may depend on his trust of Z. In many scenarios, X is given information by the system about the reputation of Y, or about the provenance of some information – it is widely accepted that these are important for trust. But even assuming that a typical X is willing to restrict his warrant for his trust in Y to reputation, provenance, recommendations and other mechanisms that have been extensively theorised online, he still needs to trust the *source* of the reputation/provenance/recommendation. If someone does not trust, say, Amazon, they are unlikely to trust the *-rating system that it hosts, even though it is intended to provide an objective assessment of Amazon's products. The provision of such information does not solve the trust problem – it just displaces it to another point of the system.

Recall also a point made earlier, that institutions can help promote well-placed trust if they are themselves trusted. It is also worth noting in this context that people contributing to a social machine, by trusting the machine's structuration of behaviour, also have to trust that their fellow users will behave in good faith. The trustworthiness of the machine will also depend on the trustworthiness of the user community. This is somewhat beyond the scope of this paper, which focuses on the challenges to designers, but the wide range of other stakeholders (owners, managers, shareholders, policymakers, users) should be an important focus of future research, and a complete social machine program should take all relevant roles into account.

5.2 The logic of trust

Z makes a claim about how Y will perform. Y in this case is the social machine, and Z the administrator. X's trust of the social machine will depend on his trust of the administrator. For instance, the motivation of the people from whom information is crowdsourced in the DARPA network challenge depended on financial incentives (a) to provide information to the administrator, and (b) to introduce new people to the group. The function of that social machine depended among many other things on enough people trusting the administration of the machine, and the likelihood of its dispersing the money.

Indeed, because we are dealing with trust with its subjective element, all that was required was that the various Xs *believed* that remuneration would be forthcoming. The money need not actually have been in place at all. Hence if we are formalising social machines using a process calculus (as advocated persuasively by Robertson and Giunchiglia), we need to make a distinction between those social properties which need to be *true* in order for a social machine to achieve its purpose, those properties which need to be *believed to be true* (but which need not be true), and those properties which need to be *both* true and believed.

This matters because a calculus should describe necessary conditions for a machine's function. In the case of the DARPA challenge, the existence of a pot of money to be distributed to the participants was *neither sufficient nor necessary* to the social machine's function. It was not sufficient, because if would-be participants were unaware of or did not believe in the financial remuneration they would not have taken part. It was not necessary, because all that mattered was that the participants were *motivated*, not that they were *paid*. Of course, this problem is most dramatic in a one-shot system, but will always re-emerge in some form even in contexts with repeat runs.

Indeed, spreading the truth about how a machine will function could on occasion undermine that very functioning. The reader may have noticed that someone helping Google by using a reCAPTCHA need not be aware that he or she is doing that (although Google makes no secret of it). This introduces an exploitative element to reCAPTCHA; one wishes to identify oneself as a human, but having done that, one is also required to perform an extra task, which is not identified as such, to help Google scan an old book.

reCAPTCHA demands very little effort, so the exploitation is probably bearable, but even so someone might resent having to help *Google* when they wanted to interact with *Facebook*. More generally, if people came to understand that, say, a social network was gathering information about them primarily in order to sell to marketing companies, or that a healthcare social machine was gleaning information primarily to sell to pharmaceutical companies, the feeling of exploitation (even if it was plausibly in the interests of the users) might have the effect

of discouraging the users from taking part. It is essential to make a distinction between what is known about the system, what users should believe (even if false) about the system, and what users should be unaware of (even if true) about the system.

5.3 Differences of interpretation

Where the interests of Z and X do not align, it is important to ensure that X's interpretation of R coincides with that of Z. This is not always the case with technology. Where Z is a designer who has created an artificial agent Y, Y's trustworthiness is often measured by Z against a highly technical specification R. However, the user X will typically see the technology holistically as part of a system with which he is confronted. If we take the example of an ID card, the system designer may be pleased to have devised a secure system. But the owner of the card will judge it in terms of the extent to which it empowers and constrains him. As Charles Raab puts it, "it is no comfort to a privacy-aware individual to be told that inaccurate, outdated, excessive and irrelevant data about her are encrypted and stored behind hacker-proof firewalls until put to use by (say) a credit-granting organization in making decisions about her" [18].

There are many types of case where R, the claim that is made about Y, can be very different from I(R,c), X's interpretation of that claim. If trust is to be maintained, R must be couched in a way that is meaningful for X. A merely technical specification of behaviour, however accurate, is unlikely to be enough. Yet a technical specification of the system's behaviour is required if we are to be able to program social machines rigorously.

6 CONCLUSION

The problem of trust is that it is hard to align to an arbitrary degree of certainty with trustworthiness. It is important, if dispiriting, to note that the most trustworthy system is useless if it is not trusted. Furthermore, it could happen that a trusted system works perfectly well (to its designers' satisfaction, anyway) *even if it is not trustworthy*.

Much will depend on the incentives given to participants. In the case of machines which provide a good user experience (for example, healthcare networking sites from which people get best practice or companionship or counselling from others with similar problems), specifying that experience will be difficult. All a designer can really specify are issues such as the privacy and security with which health data are stored. These are important factors for user trust, but the porousness of the system will also depend on the propensity of the networking humans to misuse or leak information they gain, for example from chatrooms. The nature of the user community is at least as important as the technical specification.

Taking this thought to a logical conclusion, it is likely that public trust in such machines will be highest when the public has had a say in their design and operation. The closer the relationship between trustor, designers and administrators, the better. This suggests that a focus of future research here might be the development of tools and protocols to allow communities to design social machines to their own specifications.

In machines such as reCAPTCHA and the DARPA challenge, where the humans in the loop are performing tasks subordinate to the wider goal of the system and gaining nothing intrinsic from participation, the classic trade-off of trust (that trust matters and trustworthiness is secondary, especially in one-shot games), is harder to avoid. 'Programming' of such machines using process calculi should, from the point of view of good design, make the necessary and sufficient conditions clear. Whether this promotes or restricts cynicism is an empirical question upon whose answer the future of social machines will probably rest.

ACKNOWLEDGMENTS

The work reported in this paper was funded by the EnAKTing project, EPSRC Grant EP/G008493/1. Thanks to Dave Robertson, Luc Moreau and three referees for comments.

REFERENCES

[1] A. Bernstein, M. Klein and T.W. Malone, Programming the Global Brain, *Communications of the ACM*, in press.

[2] T. Berners-Lee, *Weaving the Web: the Original Design and Ultimate Destiny of the World Wide Web*, Harper Collins, New York (1999).

[3] D. Robertson and F. Giunchiglia, Programming the Social Computer, *Philosophical Transactions of the Royal Society A*, in press.

[4] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham and M. Blum, reCAPTCHA: Human-Based Character Recognition via Web Security Measures, *Science*, 321:1465-1468 (12th Sept, 2008).

[5] K. O'Hara, A General Definition of Trust, working paper, http://eprints.ecs.soton.ac.uk/23193/, (2012).

[6] R. Hardin, Trustworthiness, *Ethics* 107:26-42, (1996).

[7] R. Hardin, *Trust*, Polity Press, Cambridge, (2006).

[8] M. Kosfeld, M. Heinrichs, P.J. Zak, U. Fischbacher and E. Fehr, Oxytocin Increases Trust in Humans, *Nature*, 435:673-676 (2nd June, 2005).

[9] S. Toulmin, *The Uses of Argument*, Cambridge University Press, Cambridge, 1958.

[10] A. Wierzbicki, *Trust and Fairness in Open, Distributed Systems*, Springer, Berlin, (2010).

[11] M. Hollis, *Trust Within Reason*, Cambridge University Press, Cambridge, (1998).

[12] A. Pentland, *Honest Signals: How They Shape Our World*, MIT Press, Cambridge MA, (2008).

[13] J. Hendler and T. Berners-Lee, From the Semantic Web to Social Machines: a Research Challenge for AI on the World Wide Web, *Artificial Intelligence* 174 156-161 (2010).

[14] M. Sloman, Policy Driven Management for Distributed Systems, *Journal of Network and Systems Management*, 2:333-360, (1994).

[15] M. Alberti, D. Daolio, P. Torrini, M. Gavanelli, E. Lamma and P. Mello, Specification and Verification of Agent Interaction Protocols in a Logic-Based System, *Proceedings of the 2004 ACM Symposium on Applied Computing (SAC '04)*, ACM Press, New York (2004), 72-78.

[16] N. Osman and D. Robertson, Dynamic Verification of Trust in Distributed Open Systems, *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India (2007), http://www.ijcai.org/papers07/Papers/IJCAI07-232.pdf.

[17] D. Kahneman, *Thinking, Fast and Slow*, Allen Lane, London, 2011.

[18] C.D. Raab, The Future of Privacy Protection, in R. Mansell and B.S. Collins (eds.), *Trust and Crime in Information Societies*, Edward Elgar Publishing, Cheltenham, (2005), 282-318.

Navigating between chaos and bureaucracy: How open-content communities are backgrounding trust

Paul B. de Laat[1]

Abstract. Many virtual communities that rely on user-generated content (social news, citizen reports, and encyclopedic entries in particular) offer unrestricted and immediate 'write access' to every contributor. It is argued that these communities do not just assume that the trust as granted by that policy is well-placed; they have developed extensive mechanisms that underpin the trust involved ('backgrounding'). These target contributors (stipulating legal terms of use and developing etiquette, both underscored by sanctions) as well as the contents contributed by them (schemes for basic quality control: patrolling for illegal and/or vandalist content, variously performed by humans and bots). Backgrounding is argued to be important since it allows avoiding bureaucratic measures that may easily cause unrest among community members and chase them away.

1 INTRODUCTION

Online communities that thrive on user-generated content come in various formats. Contents may vary considerably—from text, photographs, videos, designs and logos to source code. Furthermore, cooperation may range from 'loose' interaction: uploaded contents are presented as-is—to 'tight' interaction: an evolving product is being worked on collectively. This distinction in cooperation patterns is referred to by Dutton [1] as 'contributing 2.0' vs. 'co-creation 3.0'. Typical examples of the former are Flickr and YouTube, of the latter Wikipedia and open source software.

These communities face the dilemma of which contributors are to be accepted as members and how contributions are to be processed and published. Some communities take a cautious approach: only some categories of people are allowed to contribute, and their contributions are critically examined, by filtering before reception or moderating afterwards. A typical example is the Encyclopedia of Earth which only accepts inputs from acknowledged experts. Moreover, their appointed 'topic editors' decide who is to write the entries and who is to participate in reviewing them. In the end they have to approve of entries appearing in a public version. Other communities, though, prefer to hand out a generous invitation to their 'crowds' in order to maximize possible returns. It consists of two parts: (1) Anyone is invited to contribute content without any restrictions on entry; accordingly, access is fully open to anyone who cares to contribute; (2) Contents contributed are subsequently accepted with no questions asked and appear right away on the appropriate spot. Publication proceeds without review and without delay. In terms of Goldman [3]: no filtering is applied at the reception stage.

Which communities typically practice this two-fold institutional gesture? Let me mention some of them as far as they predominantly revolve around soliciting and reworking of *text*. I select these since it seems especially with text that the whole spectrum from contribution (2.0) to co-creation (3.0) unfolds; activities in communities which focus on other kinds of content most often remain at the level of contributing. The first category is *'social news'* sites that focus on creating a collective discussion about topics in the news that are deemed to be relevant. The formula is basically the same for all: users are invited to submit news stories and/or news links that will be put up for public discussion (comments). In this category we find Digg (2004) and Reddit (2005) which focus on news of all kinds, and Slashdot (1997) and Hacker News (2007) which focus on technology-related issues.[2] The second category is user-generated *newspapers* that have been around since 2004. NowPublic (2005), Digital Journal (2006) and GroundReport (2006) invite everybody to become a citizen journalist and contribute their own articles, blog entries and/or images to the site, as well as leave comments on those of others. These contributions essentially remain unaltered. Wikinews (started earlier, in 2004) goes one step further: in the so-called 'news room' articles which have been submitted are polished further by fellow contributors (by means of a wiki). As soon as criticisms have been met, the article can officially appear on the 'front page'. The third and final category consists of user-generated *encyclopedias.* Many such communities exist (cf. [5]), but only a few have adopted policies of open access & immediate publication. British h2g2 (2001) invites everybody to compose entries; these are put up on the site for public commenting. Wikipedia (2001) and Citizendium (2007) lean more towards co-creation by publishing new entries in an open-access wiki that allows other participants to instantaneously insert their own textual changes.[3]

2 TRUST

This gesture of unrestricted and immediate access to the community platform (to be denoted 'write access')[4] can be interpreted as a form of '*institutionalized' trust* towards prospective par-

[1] Faculty of Philosophy, University of Groningen, the Netherlands. Email: p.b.de.laat@rug.nl.

[2] Henceforth, years of foundation are given in brackets.

[3] These communities will serve as cases to be analysed further on in this article. Note that while they practiced unrestricted and immediate access from the outset, some of them recently have been pondering—or actually resorted to—more restrictive editorial policies: filtering before reception (to be commented on below).

[4] This term is in use among developers working together on open source software. As a rule, anyone may access the site and inspect the contents ('read access'). When participants have proven their skills, they may acquire the additional right to directly contribute code to a project's source code tree: they have obtained 'write access'.

ticipants. The italics are employed in order to stress two particular points. On the one hand, the gesture is an institutional one: we are dealing here with the ways in which an institution approaches the members it depends on, not with interpersonal trust. On the other hand, the gesture embodies the presumption that prospective participants are willing to contribute content with good intentions and to the best of their capabilities. Their trustworthiness in terms of moral intentions and capabilities is taken for granted. Notice that different capabilities are involved across the various communities. Social news sites rely on capabilities of argumentation and discussion; rhetoric skills are vital. Encyclopedic projects, on the other hand, are mainly interested in people's cognitive capabilities to contribute knowledge. Citizen journals occupy a position in between: they are looking for both kinds of capabilities.

That trust is at issue here can easily be seen from the fact that all communities concerned are exposing their respective repositories of content and *entrust* them as it were to the whims of the masses. They have decided to fully rely on their volunteers, thereby making themselves vulnerable and taking risks. Discussion sites, published news reports and encyclopedic entries can easily be polluted and spoiled by all kinds of disruptive actions. As Wikipedia defines the matter, 'cranks' may insert nonsense, 'flamers' and 'trolls' may enjoy fomenting trouble, 'amateurs' may ruin factual reporting, 'partisans' may smuggle in their personal opinion where this is inappropriate, and 'advertisers' may just try to promote their products anywhere (English Wikipedia:RCO). Repositories polluted in this way undermine the viability of any community, and necessitate laborious cleaning actions to be performed.

Given this gesture of fully trusting potential participants and giving them write access accordingly, which mechanisms of trusting others may be relied on in the process? Which processes possibly lie behind it? In the sequel I discuss three well-known mechanisms to handle the trust problem: the assumption, inference, and substitution of trust. Subsequently, I argue for a fourth mechanism that seems to have been neglected in the literature thus far: *backgrounding* trust. In this approach the gesture of full trust is underpinned by developing support mechanisms in the background that render the trust-as-default rule rational in a reductionist way.

First and foremost, the trust involved may be the simple *assumption* that the crowds are trustworthy. Trustworthiness is assumed without any particular evidence to support that assumption. The rationale for this assumption is that precisely by acting *as if* trust is present, one may actually produce it in the process [2]. In Luhmannian terms: the gesture of trust creates a normative pressure to respond likewise. Can any good reasons be advanced for the assumption? Which mechanism may be argued to underlie said normative pressure?

Pettit [8] argued that esteem is the driving force. Since people are sensitive to the esteem of others, they will answer an act of trust with trust as it enables them to reap the esteem that is being offered to them. As argued before [4: 332], this interpretation of the normative force of trust does not seem wholly convincing in the case of open-content communities. While esteem surely is a driving force, it would seem to be an underlying one, not a paramount one. A more forceful interpretation obtains if we move away from this calculating conception of as-if trust to another conception that is based on a vision of and hopes in the capabilities of others. As argued by McGeer [7], showing trust may be rooted in hopes to challenge others to apply their capabilities in return. These others are not manipulated but empowered to show their capacities and further develop them. The trusting party puts his/her bets on a utopian future.[5] Such reasoning can in a straightforward fashion be applied to our open-content communities since the capabilities that are the cornerstone of this McGeerian vision have quite specific connotations here. By granting unrestricted and immediate access, crowd members are challenged to show their capacities of commenting, reporting news, or contributing reliable knowledge. They are invited to fulfil the promise of a community of exciting, newsworthy, or encyclopedic content.

A second way to handle the tensions that a trusting gesture generates is to *infer* trustworthiness. One looks for indicators that inspire confidence in the other(s) as a trusted partner: perceived individual characteristics like family background, sex, or ethnicity, belonging to a shared culture, linkage(s) to respected institutions, or reputation based on performance in the past (this argument can be traced back to Zucker [11]). Moreover, the calculative balance of costs and benefits may seem to preclude a non-cooperative outcome. As argued before (in [4: 330-31]), I do not believe that an open-content community operating in cyberspace—or any virtual community for that matter—has many reliable indicators to cling to. Virtual identities are always precarious; anonymity of contributors only aggravates this problem. Even the common requirement to register and choose a user name (or even disclose one's real name) hardly alleviates the problem (cf. [5]). Moreover, contributors often just enter and leave, precluding any stable identity let alone reputation to form. To sum up: signalling trustworthiness cannot be implemented in a reliable way. So while the inference of trust has rightly been regarded a central component of processes of trust formation in *real life*, I do not think it has much value in virtual surroundings.

A third way to handle the problem of trust may be referred to as the *substitution* of trust. Wherever people interact continuously and some kind of community emerges, rules, regulations, and procedure tend to be introduced. Often these enact restrictions on behavioural possibilities. As a result, reliance on participants' wisdom and judgment in contributing is reduced; their actions become less discretionary. As a corollary, the need to grant them trust is lessened; the problem of trust is partly eliminated. The introduction of bureaucratic structure of the kind effectively substitutes for the need to estimate—or assume—participants being trustworthy. Below evidence is presented on some of our open-content communities recently instituting restrictive rules and regulations: filtering incoming content prior to publication. Write access thus becomes circumscribed and regulated.

However, a fourth mechanism to deal with the tensions of an all-out policy of trust is to be distinguished. It embodies efforts, in the absence of reliable inference, to create a middle road between relying on the normative power of trust on the one hand, and (partly) eliminating the problem by substitution on the other hand. In this approach the default rule of all-out trust is kept intact by underpinning it in the background with corrective mechanisms that contain the possible damages inflicted by malevolent

[5] McGeer uses the term 'substantial' trust, as opposed to the shallow trust Pettit is supposed to refer to. I prefer to avoid the former term since, to my view, not another type of trust is being defined, but just a different mechanism for generating trust ex post that actors may supposedly rely on ex ante.

and/or incapable contributors. To my knowledge, this approach has been neglected in the literature up to the present. As we will see, the supportive mechanisms themselves are not unknown, but their corrective function for keeping the default rule of trust intact has largely gone unnoticed.

3 BACKGROUNDING TRUST

I propose that several types of backgrounding can be distinguished (to be elaborated below in further detail). First, a cultural offensive can be launched to curb potential digressers: legal terms of use and an etiquette of sorts that defines proper behaviour are developed and propagated. Secondly, these standards of behaviour can be underscored by defining sanctions and disciplinary measures. Participants that deviate too much from the ground rules for constructive cooperation may be punished and ultimately expelled from the community. Thirdly, structural schemes can be introduced that aim to guarantee the quality of the community's contents. These range from relatively simple vandalism patrol schemes up to voting and quality enhancement programs. The bottom line for all three activities is that they may—at least partly—contribute to sustaining the rationality of the decision to maintain an editorial policy of all-out trust. They serve to keep the default rule of full trust in place.

3.1 Legal terms and etiquette

As a consequence of their full-trust write access policy, our open-content communities are quite vulnerable to disruptive behaviour, from posting illegal content to vandalist actions. As a way of defence they are first of all trying to lay down legal guidelines. Plagiary, libel, defamation, illegal content and the like are strictly forbidden. This is considered the baseline for proper behaviour since deviations from them would land the site with legal trouble.

Interestingly, though, our communities under study also promote 'good manners' *beyond* these legal terms of use. An etiquette is formulated for regulating mutual interactions on their sites. Leaving Wikinews and Wikipedia aside for the moment (see below), all of them stress the same kind of exhortations in their 'community guidelines', 'house rules', 'netiquette', or 'rediquette'—be it to varying degrees.[6] On the positive side, members are urged to always remain respectful, polite, and civil; to stay calm; to be patient, tolerant, and forgiving; to behave responsibly; and/or to stay on topic at all times. On the negative side, the list of interdictions is much longer. One is urged to refrain from calling names, offensive language, harassment, and hate speech. Flaming and trolling are sharply condemned. Commercial spam and advertisements are declared out of bounds. Flooding a site with materials that are offensive, objectionable, misleading, or simply false only amount to an objectionable waste of the site's resources (nicknamed 'crapflooding').

Finally, let us consider Wikinews and Wikipedia. Both under the umbrella of the Wikimedia Foundation, they have adopted virtually the same etiquette (called: Wikiquette). It is in fact the most extended set of rules for polite behaviour in open-content communities to be found anywhere on the Net. Assuming good faith on the part of others—and showing it yourself— is the starting point. Help others in correcting their mistakes and always work towards agreement. Remain civil and polite at all times: discuss and argue, instead of insulting, harassing or personally attacking people. Be open and warm. Give praise, and forgive and forget where necessary. Overall, several pages are devoted to the subject (http://en.wikinews.org/wiki/Wikinews:-Etiquette; http://en.wikipedia.org/wiki/Wikipedia:Wikiquette).

3.2 Enforcement

Both legal rules and etiquette cannot do without some mechanism of enforcement. With all communities above, without exception, sanctioning of deviant users has become the normal state of affairs. Users that (repeatedly) flout the rules of etiquette—let alone the legal rules—can be banned from the community for some period of time, or even forever. As a rule the professional editors as employed by the site ('editorial team') simply assume these judicial powers themselves. With others, site volunteers are entrusted with the task. At h2g2, these are appointed for the job (as 'moderators') by staff of the company which owns the site (formerly the BBC). The pair of Wikipedia and Wikinews appoints candidates with a procedure that relies on public consultation of the community ('administrators'). Citizendium does likewise ('constables').

The mechanisms of rules & sanctions taken together send the message: respect legal terms of use and be civil and polite—otherwise thou risk to be expelled. Notice how these may impact on the employed policy of unrestricted and immediate access. That policy assumes trustworthiness of participants from the outset. Inculcating respect for legal issues and rules of etiquette then may serve to *create* trustworthiness where it is found to be lacking—*afterwards*. Whenever the assumption of trustworthiness appears unwarranted, that defect can (at least partly) be repaired afterwards. As a result, the full write-access policy is underpinned and can possibly remain in force after all. 'Backgrounding', as I shall call this phenomenon, keeps confidence in full-trust as the default intact.

I would argue, however, that these mechanisms can do just so much. They can only possibly 'educate' participants that are staying longer. Newcomers, who are the most likely source of mischief, can hardly be supposed to have read let alone internalized the rules involved upon entry. As a result, the campaign for legal and civil conscience has no effect on them, and the full-trust policy remains vulnerable to their abuse. Therefore we now turn to structural means that may support the full-trust policy. No longer the dispositions of people but the contents they actually contribute come in focus. I shall argue that these tools are ultimately able to do a more powerful job of sustaining that policy.

3.3 Quality management

The term 'quality management' is used in quite a broad meaning: it is to refer to both *rating* and (for dynamic entries) *raising* the quality of contributed content, throughout the whole quality range, from low to high. At the lower end, the mess of clearly inappropriate content that flouts basic legal terms of use or etiquette has to be cleaned up. Beyond these tasks of 'basic cleaning' (as I shall label them) the quality of content—as far as it has passed the former test of scrutiny—can be monitored continu-

[6] For reasons of space, precise references to the various community sites are omitted (but are available on request).

ously and (in case of dynamic content) raised ever further. Such quality schemes may already be the normal modus operandi (cf. the wiki format); they may also be developed as additional mechanisms since the basic mode is felt to be an insufficient guarantee for quality.

3.3.1 Social news sites and citizen journals

Social news sites and citizen journals (apart from Wikinews) are usefully treated together since all operate in the 'contributing 2.0' mode. These solicit stories (whether existing—for social news sites, or newly composed—for citizen journals) and comments on them. Tasks of basic cleaning are performed (afterwards) by the editorial teams involved: they scout their sites continuously for illegal and inappropriate content. Usually, site visitors are also solicited to report 'violations'. Any content of the kind—whether illegal content, flooding, spamming, advertising, hate speech or abusive language—is immediately dealt with and deleted; those who posted them are reprimanded or, after repeated violations, banned from the site.[7] Such basic cleaning can however just achieve so much: the quality of contents *above* the baseline of appropriate content remains an issue.

In order to tackle this thornier problem these sites have pioneered a novel approach: stories and comments can be *voted on*, usually as either a plus or a minus. As a rule, all users are entitled to vote. Note though that some communities require registration, and in Slashdot the right to vote obtains for a limited amount of time only. Let me elaborate these schemes. Digg has pioneered 'digging': if a user 'likes' the content, it is digged (+1), if (s)he 'dislikes' it, it is buried (-1). GroundReport has adopted the very same scheme. Reddit, Hacker News, and Slashdot use the more neutral wording of voting for the process: a plus if entries are found to be 'helpful', 'interesting', or 'constructive', a minus if they are not. Finally, NowPublic and Digital Journal only allow plus votes, for articles deemed 'newsworthy'.

The sum total of votes then determines the *prominence* of articles on the site. By default, stories (on the front page) and comments on them (below each story) are displayed in chronological order of submission, with the most recent ones on top. Entries thus have a natural rate of decay. Voting data, fed into one algorithm or another, then force the liked items to remain longer on top of the page (countering natural decay), while at the same time forcing the disliked items—at least as far as 'dislikes' are part of the scheme—to plunge down the page quicker (accelerating natural decay).[8] Slashdot uses a slight variation: with vote totals for items being limited to the range -1 to + 5, readers can choose their own personal threshold level to determine whether items become *visible* to them or not when they enter the site. Thus articles of bad repute are no longer punished by being pushed down the page, but by being 'deleted' for all practical purposes.

3.3.2 Encyclopedias and Wikinews

The remaining communities in my sample operate in proper 'co-creation 3.0' mode (Wikinews and encyclopedias). They also resort to basic cleaning concerning illegal or inappropriate content; in addition they have introduced elaborate quality schemes that go beyond simple voting. Let me start with h2g2 that does not use the wiki format, but just old-fashioned commenting. Tasks of basic cleaning are executed by the aforementioned volunteer 'moderators' (as appointed by the owner). As they phrase it, someone has to 'clean the flotsam'. In addition, these decide on banning users who are found to be in violation. Higher up the quality scale, authors may strive for their article to appear in the 'edited guide'. To that end, it has to be put up for public review, be recommended by a 'scout', and edited by 'subeditors'. Notice that these two roles (volunteer roles one has to apply for) are intended to support authors, as opposed to control them. They are urged to operate as 'first among equals'.

Citizendium, Wikipedia, and Wikinews have the wiki mode of production in common. This wiki is *the* place to carry out basic cleaning of illegal and inappropriate contents. Users are always on the alert regarding contents, allowed to immediately correct new edits in the wiki, and invited to 'report' any transgressor to the authorities concerned (constables and administrators respectively). The three communities have quite similar procedures as well for identifying and promoting high quality content (apart from normal 'wikiing'). In Citizendium an entry may gain the status of 'approved'. To that end, an appointed moderator (denoted 'editor') has to give his/her approval. This role incumbent is also to exercise 'gentle oversight' concerning matters of evolving content. So here again, like in h2g2, a non-authoritarian role, a 'primus inter pares'. Wikinews and Wikipedia, on their part, elaborated wholly public procedures for entries to gain the status of 'good' or even 'featured' article. As a preliminary step towards acquiring such statuses an entry may be put up for public 'peer review' first.

Wikipedia in particular, though, over time has come to developing *additional* efforts of quality management that supplement the basic wiki mode of production. The most extended quality-watch program anywhere in our communities is to be found here. It revolves around a kind of permanent mobilisation of Wikipedians who are invited to focus their energies on quality enhancement. In their fight against 'vandalism' basic cleaning is high on the agenda. Users can maintain personal 'watch lists': listed entries are kept under surveillance for new edits coming in. 'New Pages Patrol' is a system for users to scan newly created entries for potential problems right after they are submitted. Furthermore hundreds of software bots have been developed for the purpose. After severe testing and public discussion within the Wikipedian community, these may be 'let loose' on a 24 hours basis. A famous example is Cluebot, which is instructed to intervene whenever suspicious words are inserted ('black lists') or whole pages deleted (http://www.acm.uiuc.edu/~carter11/-ClueBot.pdf). The 'new generation' CluebotNG operates along quite different lines: as a neural network. The bot has to be fed with both constructive and vandalist edits. By interpreting those data it hopefully will learn in the long run to correctly diagnose instances of vandalism (http://en.wikipedia.org/wiki/User:-ClueBot_NG).

Close watch also extends beyond the issue of vandalism. Wikipedian pages and articles are under constant surveillance whether they should be kept, deleted, merged, redirected, or 'transwikied' (=transferred to another Wikimedia project). More im-

[7] In Reddit, those who started a 'subreddit' usually are awarded the same powers for their particular subreddit.

[8] Some basics of these algorithms are elaborated in http://www.seomoz.-org/blog/reddit-stumbleupon-delicious-and-hacker-news-algorithms-exposed.

portantly, in order to raise the quality of entries further, 'WikiProjects' (with subordinate 'taskforces') are formed in which people focus on specific themes (such as classical music or Australia). Each project takes relevant entries under its wings and promotes improvement. In particular they are entrusted the task of grading the articles in their purview by quality (7 degrees, the highest being featured and good, cf. above) and importance (4 degrees) (http://en.wikipedia.org/wiki/Wikipedia:WikiProject_-Council/Guide). Last but not least, tools are made available to users for judging the credibility of entries: the WikiTrust extension and the WikiDashboard. These tools calculate proxies for credibility of entries from their review histories. Users may use these indicators for focused quality enhancement of entries.

3.3.3 Intensity

Before embarking on a discussion of the relationship between measures of quality control and trust, let me first put them in a comparative perspective across the whole range of open-content communities under study. Legal rules and etiquette (3.1 and 3.2) seem to be emphasized throughout, in about equal measure. This stands to reason, since these revolve around behavioural norms of trust and respect which are universally applicable to all communities of open textual content. Not so however for quality management efforts: these are clearly intensifying if we move towards the encyclopedic end of the range. For one thing, patrolling for improper content is increasing. For another, voting schemes make way for a variety of teams that focus on quality within the wiki mode. Why this more intense mobilisation?

I want to argue that this is mainly due to the different types of content involved. Social news sites aim to foster discussions; an exciting exchange of *opinions* is what they are after. These discussions, moreover, have a kind of *topicality*—in the long run their importance simply fades away. To that end, a 'contributing 2.0' mode is sufficient. In order to guarantee quality in this mode, scouting for inappropriate content combined with voting schemes is good enough: good discussions will remain in view (longer), while bad discussions will disappear out of sight (quicker). The natural tendency for time to produce 'decay' is intensified. To citizen journals, furthermore, similar arguments apply.

Encyclopedias, however, aim to render the *'facts'* about particular matters. Such entries cannot be produced in one go, but have to evolve over time. Moreover, such entries are to remain *permanently* visible, ready to be consulted. For the purpose, 'co-creation 3.0' is the preferred mode: Wikipedia, Wikinews, and Citizendium have chosen the interactive wiki format as mode of production (which does not necessarily have to be so: h2g2 prefers a 'contributing' approach). Obviously, such a dynamic mode is susceptible to disruptions. Watching over quality therefore becomes a more urgent task. For that purpose, the wiki is turned into a space of intense patrolling and quality enhancement efforts.

3.3.4 Backgrounding trust

After this assessment of quality management efforts across our sample of open-content communities finally their connection with the default rule of full trust concerning write access remains to be specified. To what extent may this institutionalized trust be said to be 'backgrounded' by quality control? As far as this control is concerned with basic cleaning tasks, there *is* a connection. Scouting for inappropriate or outright vandalist contributions—whether inside a wiki or not, whether by special volunteer patrol teams or the editorial team only, whether by humans or bots—, combined with appropriate corrective action and disciplining of transgressors, is a contribution to keep the policy of full write access viable. Since disruptive contributions can always be sifted out afterwards, the gates may remain open to all. 'Backgrounding' of the kind may effectively allow unrestricted and immediate write access to remain the default.

All other efforts under the rubric of quality control—which push for quality promotion—are *not* connected to trust: voting schemes in order to push high quality articles to prominent and/or visible position (social news and citizen journals), efforts to promote articles to the 'edited guide' (h2g2), to develop 'approved' articles (Citizendium), or to produce 'good' or 'featured' articles (Wikinews, Wikipedia) hardly bear a relationship. Though profiting largely from the condition of full write access for everybody since a maximum of contributions is being solicited, these ongoing initiatives obviously cannot be considered to support—or undermine for that matter—the institutional trust exhibited. They just thrive on it.

4 DISCUSSION

As regards quality management (3.3) critics may object that the relevant rules, regulations, and procedures cannot neatly be sorted into those that either background or substitute trust (or are neutral in that respect); they are just variations on the same theme of concern for quality that only differ in their temporality of application. I would argue, however, that the distinction is sound and important. My argument proceeds along the following lines.

On the one hand, schemes for quality control can aim directly at the discretion of participants and reduce it (e.g., filtering). This reduction of discretion by definition leaves less-than-full-trust to participants. As a corollary, hierarchical distinctions among participants need to be defined (such as determining who is entitled to carry out filtering, and who is to be subjected to it).[9] If so, some amount of bureaucracy proper has been introduced in the community. Note finally, that the substitution of trust as effectuated is precisely the intention of such schemes. On the other hand, measures of quality control can also buttress policies of write access for all (e.g., scouting and patrolling for vandalism, whether by humans or bots). Institutionalized full trust remains a viable option because of the 'damage repair options' that are unfolding. Essentially these schemes mobilize the whole community—and therefore do not introduce any hierarchical distinctions. Furthermore, the supporting effect on institutionalized trust towards participants is more properly a side effect; the main focus of such campaigns is quality overall. Obviously, in between the two categories quality management initiatives can be discerned that do *not* touch upon our issue of institutional trust. The above mentioned voting and quality rating schemes are cases in point.

The contrast can best be captured in terms of the trust assumptions embodied in the various write access policies involved. In the case of patrolling new inputs and new contributors (as well as quality watch and voting schemes more generally), the assumption of full trust of potential participants is left intact and

[9] Cf. by way of analogy the common distinction between developers and observers in open source software projects.

untouched. The default remains: 'we trust your inputs, unless proven otherwise.' In the case of filtering which reduces the trust offered, this default is exchanged for quite another one: 'we can no longer afford to trust your inputs, and accordingly first have to check them carefully.'

In line with the above I want to underline that backgrounding trust in open-content communities is very important for their functioning. The mechanism allows the full-trust write access policy to remain in force. By the same token, other available mechanisms to manage the trust problem do *not* have to be resorted to. In particular, the substitution of trust by installing bureaucratic measures can be avoided. Before elaborating this point let me first provide some examples of steps towards bureaucracy as considered or actually taken by our communities.[10] The Slashdot editorial team routinely scans incoming stories and only accepts the 'most interesting, timely, and relevant' ones for posting to the homepage. Furthermore, since 2009, Now Public and GroundReport filter incoming news before publication. With the former, first articles from aspiring journalists are thoroughly checked by the editorial team; subsequent ones may go live immediately and are only checked afterwards. With the latter, the site's editors have to give their approval to all proposed articles prior to publication. Only reporters with a 'strong track record' have full write access. In the Wikimedia circuit, finally, proposals for checking incoming edits for vandalism before publication have been circulating for several years; only after approval edits are to become publicly visible. Such review is to be carried out by experienced users. In this fashion, evidently, trust in newcomers gets restricted. The proposal is actually in force in a number of their projects from 2008 onwards: Wikipedia and Wiktionary (German versions), as well as Wikinews and Wikibooks (English versions).[11] [12]

Why then would it be important to avoid bureaucracy? The answer is that such measures may meet a chilly reception and cause unrest and trouble among community members. A conspicuous example of such unrest is the heavy contestation of the system of reviewing edits prior to publication (called 'Flagged Revisions') in English Wikipedia: the proposal has encountered fierce resistance and finally had to be abandoned (cf. [6]). Community members may simply detest bureaucratic rules and threaten to withdraw their commitment accordingly. That is why backgrounding trust is such an important mechanism.[13] Note also in this context the conspicuous role of software bots in Wikipedia. These have been and still are very active in detecting vandalism—often ahead of patrollers of flesh and blood. The home page of Cluebot is full of 'barn stars' from co-Wikipedians, awarded since the bot had detected vandalist edits before them, in just a few seconds. Reportedly it identifies, overall, about one vandalist edit per minute (over a thousand per day). Thanks to Cluebot and its likes, introduction of the system of Flagged Revisions was not inevitable and the plans could be shelved.

Recently both Simon [9] and Tollefsen [10] asked themselves the question: can users rely on Wikipedia? In their affirmative answers they pointed to editorial mechanisms in place that may ensure *high quality*: the wiki format with associated talk pages [9: 348], and the procedure for acquiring 'good' or 'featured' status [10: 22]. My question has been a slightly different one: can Wikipedia trust their users and grant them unrestricted and immediate write access? No wonder my—equally affirmative—answer turned out to be slightly different. Contributors can fully be trusted since swift procedures to filter *low quality* submissions afterwards are in place; in complementary fashion, a continuous campaign among participants promotes respect for etiquette and basic rules of law.

[10] For reasons of space, references that document the steps to be mentioned have been omitted—but are available from the author.

[11] The proposal is also in force in several smaller language versions other than English, German, or French (cf. http://meta.wikimedia.org/wiki/-Flagged_Revisions).

[12] In our sample it is editorial teams (social news sites, citizen journals), moderators (h2g2), constables (Citizendium) and administrators (Wikipedia, Wikinews) who hold the powers to clean up messy content and/or to discipline members. Obviously, these power holders also represent bureaucracy—the difference with the filtering measures mentioned being, that no community members seem to be opposed to such a baseline of bureaucracy.

[13] Note in this respect how some of our communities try to bolster the quality process by introducing specific supportive roles that are intended as 'prime among equals' (cf. 'editors' in Citizendium, and 'subeditors' in h2g2). Their intention is clearly to *avoid* introducing hierarchical relations in this fashion. But trying to operate as such a 'primus' is walking a tight rope: in his/her performance, the role occupant may easily come to be perceived as an ordinary boss.

REFERENCES

[1] W.H. Dutton. The wisdom of collaborative network organizations: Capturing the value of networked individuals. *Prometheus*, 26(3): 211-230 (2008).

[2] D. Gambetta. Can we trust trust? In D. Gambetta (ed.), *Trust: Making and breaking cooperative relations*, Blackwell: Oxford, 213-237 (1988).

[3] A.I. Goldman. The social epistemology of blogging. In J. van den Hoven, J. Weckert (eds.), *Information Technology and Moral Philosophy*, Cambridge University Press: Cambridge etc., 111-22 (2008).

[4] P.B. de Laat. How can contributors to open-source communities be trusted? On the assumption, inference, and substitution of trust. *Ethics and Information Technology*, 12(4): 327-341 (2010).

[5] P.B. de Laat. Open source production of encyclopedias: Editorial policies at the intersection of organizational and epistemological trust. *Social Epistemology*, 26(1): 71-103 (2012).

[6] P.B. de Laat. Coercion or empowerment? Moderation of content in Wikipedia as 'essentially contested' bureaucratic rules. *Ethics and Information Technology*, 14(2): 123-135 (2012).

[7] V. McGeer. Trust, hope and empowerment. *Australasian Journal of Philosophy*, 86(2): 237-254 (2008).

[8] Ph. Pettit. The cunning of trust. *Philosophy and Public Affairs*, 24(3): 202-225 (1995).

[9] J. Simon. The entanglement of trust and knowledge on the Web. *Ethics and Information Technology*, 12(4): 343-355 (2010).

[10] D.P. Tollefsen. Wikipedia and the epistemology of testimony. *Episteme*, 6(1): 8-24 (2009).

[11] L.G. Zucker. Production of trust: Institutional sources of economic structure, 1840-1920. *Research in Organizational Behaviour*, 8: 53-111 (1986).

Artificial and Autonomous: A Person?

Migle Laukyte*

Abstract. Autonomy and personhood are two statuses the law usually ascribes to human beings. But we also ascribe these statuses to nonhuman entities, notably corporations. In this paper I explore the idea of expanding this ascription so as to include a *third* class of entities: not only humans and corporations but also artificially intelligent beings (artificial agents). I discuss in particular what autonomy and personhood mean, and I consider different ways in which these statuses can be applied to artificial agents, arguing that although computer science and software engineering have yet to develop such agents, a circumstance that makes the whole discussion hypothetical, it still makes sense to discuss these issues, on the assumption that once the former status (autonomy) is built into these agents, the latter status (personhood) will become a more realistic scenario.

1 INTRODUCTION

Individual autonomy and legal personhood are two interrelated notions: once a human being achieves full autonomy as an adult, that person becomes a subject of rights and duties, that is, he or she becomes a person in the eyes of the law.

Autonomy and personhood, however, are not something the law ascribes exclusively to humans: we have extended these statuses to nonhuman entities as well, such as corporations, ships, and other artificial legal persons.[1]

This paper revolves around the idea that our ascription of autonomy and legal personhood may be still in process, specifically as concerns artificially intelligent entities (from here on out "artificial agents"), which I posit as a third class (next to humans and corporations) to which these two statuses may be ascribed.

The paper is divided into two main parts: the first deals with autonomy, which I take to be an essential requisite of artificial agents before any personhood can be ascribed to them. Autonomy is discussed as both a philosophical and a computational concept, and in both respects I will be attempting to determine what it takes for an artificial agent to be autonomous.

The second part of this paper will thus turn to the issue of legal personhood, asking whether artificial agents should be recognized as persons once they become fully autonomous in both the philosophical and the computational senses I will be clarifying. In fact, one can easily envision the consequences that might accompany the development of artificially autonomous agents, and since these are too broad to be discussed intelligibly in the space of a single paper, I will restrict my discussion to what such a development would entail for the law. I speculate that we would have to revisit the concept of legal personhood as a status acquired in consequence of gaining autonomy. I also discuss in this connection the question of whether autonomous artificial agents should be likened to natural persons (humans), or to artificial ones (corporations), or whether we should work out a new formula for such entities.

The paper is thus organized as follows: in Section 2, I introduce a Kantian concept of autonomy as self-governance. I then apply this concept to artificial agents, asking whether this is a useful basis on which to proceed in building agents. I argue that this is not a possibility given the current state of the art in computer science (CS), and I therefore suggest that we focus on the concept of autonomy adopted in CS itself: Section 3 discusses how this concept can be applied to artificial agents. Then, in Section 4, I consider what the development of artificial autonomous agents would mean for the law. I argue in particular that if an agent is autonomous, it is responsible for its actions, and only legal persons—natural ones (people) or artificial ones (corporations)—are held responsible for their actions in law, and the question becomes which of these two classes is the more appropriate basis on which to consider the responsibility of an agent as a legal person. Sections 5 and 6 discuss these two hypotheses, respectively, and Section 7 puts forward a few ideas about how we could deal with these issues going forward.

2 KANT, AUTONOMY, AND ARTIFICIAL AGENTS

In this section I present a concept of autonomy based on the account of it that Kant expounds in [1], and the reason why I look to Kant is that his account lays the modern foundation of the concept and is often taken as the starting point in understanding the idea of autonomy and working out its implications in different settings.

Kant introduced what in his time was a revolutionary conception of morality [2], which he called self-governance or autonomy, arguing that such autonomy lies in the will: "The autonomy of the will is the sole principle of all moral laws, and of all duties which conform to them" [1].

What Kant meant is that in order for someone to be recognized as a moral agent, he or she must be a self-governing, or autonomous, creature. Which in turn means that we are the makers of our own action: we are self-legislating creatures who follow their own moral law, and a failure to do so is a failure on our part to act as moral agents. Thus Kant considered autonomy a compass that enables "common human reason" to tell what is consistent with duty and what is not (or what is moral and what is not). This "common human reason," or pure practical reason, belongs to all of us: this is why we can understand and relate to

* CIRSFID, Bologna University School of Law, via Galliera 3, 40121 Bologna. Email: migle.laukyte@unibo.it.

[1] As was briefly hinted at a moment ago, an artificial person can take different forms aside from the aforementioned corporations: states and municipalities, for example, can also be so considered. Still, for the sake of expediency, I will be taking the corporation (a business entity having a separate existence from its owners and managers) as a paradigmatic example of what an artificial person is.

one another; and since we are all anchored to it, we cannot loose moral capacities no matter how corrupt we may become, because "the commonest intelligence can easily and without hesitation see what, on the principle of autonomy of the will, requires to be done" [1].[2]

This is a very simplified idea of Kantian autonomy, but even in this stripped-down version we have enough to go on in deciding whether, and if do how, autonomy so conceived is an attribute we can ascribe to artificial agents. This is a question we ask because autonomy as a philosophical concept is inherently bound up with freedom, will, and morality—three attributes that are assumed to be distinctively human. So, how can artificial agents become autonomous in the sense described? This question I will I will try to answer in what follows.

To begin with, the idea of morality as an exclusively human property is no longer an axiom. It is argued in [3] that artificial agents can take part in moral situations, "for they can be conceived of as moral entities (as entities that can be acted upon for good or evil) and also as moral agents (as entities that can perform actions, again for good or evil)." Furthermore, according to [4], if we are working to develop autonomous agents, we have to make them moral, that is, we have to equip them with "enough intelligence to access the effects of their actions on sentient beings and act accordingly," while [5] sees agents as having moral virtues—grouped in into altruistic ones (such as non-maleficence and obedience) and egoistic ones (such as self-protection)—and claims that these are the virtues we should build into agents.

It is not questioned in [6] whether artificial moral agents will be among us, and so the discussion instead focuses on what their development toward a full morality might look like: first, agents acquire moral *significance*, in that they can make decisions pregnant with moral meaning; then they acquire moral *intelligence*, in that they can reason on the basis of value and principle; and third, agents are able to learn from their moral experience, thereby acquiring *dynamic* moral intelligence, and only then can they become *fully moral* agents, when their dynamic moral intelligence further makes them conscious, self-aware, sensible, deliberative, and capable of introspection, at which point they would be recognized as having personal rights and duties (as will be shortly discussed).

So, if we assume that moral agents are possible, then the next question is: How can moral values be built into artificial agents, considering that an agent's capacity to act in accordance with moral values is inseparable from the agent's autonomy in the Kantian sense?

Three approaches are offered in [4] in working toward this goal. The first is to program directly into the agent the values the agent should be guided by, but this is quite problematic because, for one thing, we have to decide on a set of values by which an agent is to be guided, and, second, it is not quite clear what the algorithm would have to be like for each of these values, especially considering that we do not have an agreed view of what they each mean: How is an honest agent supposed to act? Can two responses to the same problem or situation be equally honest? And isn't honesty (along with any other moral trait) to be judged by an agent's action as much as by its *reasons* for action?[3]

The second approach is to make agents moral by associative learning, that is, by having them adopt the techniques by which the children learn what is morally acceptable and what is not. But the problem is that the children learn what is good and what is bad because someone explains or shows them *why* something is good or bad. This means that children learn to distinguish the good from the bad by virtue of a desire to avoid punishment or to gain the approval of their parents or the acceptance of other children. In order to learn the way children do, artificial agents should also have *motives* for action, but is that possible?

There is also a third approach, which consists in simulating the evolution of agents. The underlying idea in this case is that the agent is moral if it is rational in the sense involved in the game of the iterated prisoner's dilemma (PD).[4]

The iterated PD differs from the simple PD by virtue of its being played more than once: players do not know how many iterations there will be, but they remember each other's previous actions and will model their strategy of future actions by taking this information into account. We can find examples of this situation in nature, for it has been shown that "organisms which have mutually iterated PD interactions evolve into a stable set of cooperative interactions" [4] based on survival values. In the agent-based scenario, these survival values could take the form of moral rules.

Thus agents should cooperate and behave in a morally acceptable way. But the problem with this approach is that human morality is much more complex than what the PD can account for, and if we want to frame our interactions in game-theoretical terms, the PD framework is only one option and not even the best one [5]. It is argued in [4] that what agents need is an ability to construct a conception of morality. This is an ability we humans have, but which CS is far from being able to model. Human morality is a much and long debated concept which for this reason cannot be contained within any single *conception* of morality, or any single view of what morality is and requires of us. Indeed, the very idea of morality as a source of requirements or imperatives may not be so straightforward as it might at first blush appear, if we only take into account the connection that morality has been found to bear to the emotions—consider Hume's idea that "moral distinctions are derived from the moral sentiments" [26], such as empathy—since the emotions have a phenomenological quality to which we cannot strictly ascribe the moral properties necessary for them to count as inherently normative. In addition, many ingredients go into mortality that do not appear to be susceptible of artificial modelling: some of them are substantive, such our upbringing and the conventions forming our social milieu; others are formal, consisting of capacities that can take any range of contents, such as the capacity to "adopt personal projects, develop relationships and accept commitments to causes, through which [our] personal integrity and sense of dignity and self-respect are made concrete" [24]. So the point is that it would be quite a challenge to pack all this material into a single, comprehensive yet coherent account of moral action: we cannot do so as an

[2] Kant is not the only philosopher who thought of autonomy as strictly related to morality: also in the same line of thought were Nietzsche, Kierkegaard, Popper, and Sartre, among others. For an overview, see [7].

[3] It should be pointed out, however, that research in machine ethics has become a field of study in its own right (see, for example, [8]).

[4] A discussion of the prisoner's dilemma can be found in the *Stanford Encyclopedia of Philosophy* at http://plato.stanford.edu/entries/prisoner-dilemma/.

academic exercise, much less as an implementation of CS technology.

Hence, we can see that it is at present a task too complex for CS and software engineering to model a moral or autonomous agent in Kant's sense. Therefore, for the time being we have to set aside the Kantian conception of autonomy as moral self-governance and consider another conception of autonomy, that is, autonomy as understood and used in CS.

3 AUTONOMY IN COMPUTER SCIENCE AND ARTIFICIAL AGENTS

The main difference between the idea of autonomy in CS and the same idea in other fields of study (such as law, economics, and philosophy) is that in CS this idea is quite loose: as is argued in [9], autonomy is a widely used term in artificial intelligence, robotics, and other related fields in CS, but at the same time it is not clear what distinguishes autonomy from non-autonomy, nor is there a single pattern that can be recognized in its different uses. The result is that, while in other areas of study one can tell with relative ease when an action is autonomous and when it is not, in CS this distinction is not so clear, especially as concerns artificial agents.

We ask: What is to be considered an autonomous entity or tool in CS? The qualifier *autonomous* is applied to mobile robots, for example, and to systems and devices that show some level of intelligence or independent control ([10], [11], [12]), but none of these devices, systems, or entities can be considered *fully* autonomous, because their autonomy is a matter of subjective evaluation: what counts as autonomous action for one computer scientist doesn't for another.

So, when can we say that an agent is autonomous? There are different views in this regard, but computer scientists mostly agree that an agent is autonomous if it can (*i*) learn from experience and act (*ii*) over the long course (*iii*) without the direct control of humans or of other agents. Let us take a closer look at these three aspects of autonomous action.

The first aspect, identifying an ability to learn from experience, entails an agent's ability to accordingly modify its programmed instructions and develop new ones [4]. Hence, the more it learns, the more it will become autonomous. This is a naturalistic account of an agent's autonomy: animals are born with this knowledge, and that enables them to survive. It has been suggested, in [13], that the same can be said of agents. In fact, [14] identifies learning as one of the current trends in autonomous robotics, meaning that the focus has shifted from an emphasis on movement to one on cognition and learning.

The second aspect identifies an agent's ability to act autonomously in its environment over time [15]. Autonomy in this sense has no temporal limits, in that no agent can be considered autonomous if its instructions either "run out" or commit the agent to repeating the same pattern of action over and again.

The third aspect, identifying an ability to act without the direct input of humans or of other agents, means that an autonomous agent can control its own actions and internal states [16]. The idea of such twofold autonomy is also expressed in [17], describing autonomy [16] as being in the first place *unpredictable*, with its freedom from human intervention, and in the second place as *dynamic*, with its control over an agent's own actions (see Figure 1 below).

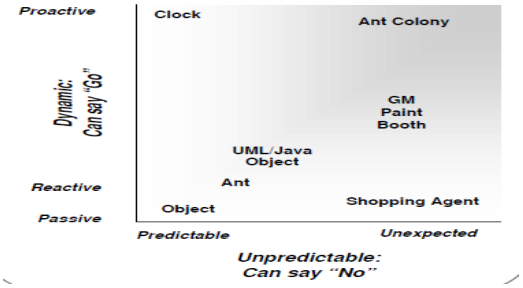


Figure 1. The concept of autonomy

Figure 1 gives an illustration of what such twofold autonomy means in a shopping agent: the agent possesses unpredictable autonomy (for it controls its own actions), but is not autonomous in a dynamic sense (humans do intervene to make it act). My idea of an autonomous agent would locate it further down on the scale of unpredictability and dynamicity, somewhere close to human action. On this view, an autonomous agent would have to be free from human intervention and would have to control its own actions and internal states.

This latter agent, in other words, should possess what [18] calls internal autonomy, or autonomy in the strong sense of the term, meaning an ability to choose not only the means to achieve goals but the goals themselves: autonomy in a weak sense means that an agent can only choose among alternative ways of achieving a predetermined end set by someone other than the agent itself; only when an agent can choose both the means and the end can it be described as autonomous in a strong sense, with characteristics essentially equivalent to those which typify what [24] calls a significant autonomous entity, one that "can shape [its] life and determine its course." Such internal or significant autonomy is also crucial to the concept of legal personhood. In fact, we humans are legal persons because we can make choices on our own and act accordingly.[5]

Let us consider how such an agent would look like in practical terms. Imagine the agent in question is an online travel agent that can "hear" me saying that my dream is to go to New York for Christmas, and that, "motivated by friendliness and social convention" [18], decides to give me a gift. Such an agent would be conscious and could be considered as an "imitation of life" [14]: it would share with us emotions such as friendliness and social inclinations, and so would be closer to the human world and distant from the world of automata.[6]

For the time being, however, CS has not yet advanced to the point of giving us such fully autonomous artificial agents. Even

[5] The same conception can be appreciated in the law's consideration of corporations as legal persons: a corporation cannot be so considered unless it is assumed to be able to make choices and act on them as a person does.

[6] In fact, scientists (see, for example, [19]) have begun to pay more and more attention to the importance of emotions in the mechanics of rational thinking: "If we want computers to be genuinely intelligent, to adapt to us, and to interact naturally with us, then they will need the ability to recognize and express emotions, to have emotions, and to have what has been called 'emotional intelligence.'" Emotional intelligence would thus lead to autonomous software agents in the most human sense of the term.

so, this should not be taken to mean that we ought not concern ourselves with the question of what would happen if such agents were with us, because we can all agree that this scenario, however much removed from the present it may be, is not thereby fantastical but is rather a concrete prospect. Hence, in what follows, I discuss this latter possibility from the legal point of view, arguing that the first legal concept we will need to reconsider when such agents will be built is that of legal personhood.

4 ARTIFICIAL AUTONOMOUS AGENTS AS LEGAL PERSONS

I consider in this section what a full and complete, human-like autonomy of artificial agents would mean for the law: if agents are *fully* autonomous, then they must be aware of their actions. If they are aware of their actions, then they must also be held to account; that is, they are liable for their actions. An agent's autonomy in law, in other words, means that the agent has rights and a corresponding set of duties. In law, rights and duties are attributed to legal persons, both natural (such as humans) and artificial (such as corporations). Therefore, the moment we deem artificial autonomous agents liable for their actions, we ascribe legal personhood to them.

If that should happen, artificially autonomous agents would have to come to be part of the class of legal persons, and the law would then have to reconsider the existing concept of legal personhood and decide whether the current legal system is adequate for the new reality, and how it should otherwise be reshaped so as to enable it to include the new artificial entities.

If we want to see whether the concept of legal personhood currently in use can account for artificially autonomous agents, we have to look at what types of legal persons exist, and whether a parallel can be drawn between existing legal persons and an autonomous artificial agent.

The concept of legal personhood has evolved over time: it is in a sense coextensive with human moral development, in that its range of application has expanded in proportion as our "social likings" have also done so, meaning that, on this ideal evolutionary line, we first extended such likings to those around us, then to the community, then to the races, then to handicapped, and finally to animals [20]. Furthermore, the concept of legal personhood has evolved in parallel with moral and political conceptions of personhood, where from ancient times the person represented "someone who can take part in, or who can play a role in, social life, and hence exercise and respect various rights and duties" [25]. Modern democracy has attributed further moral powers to the person (the capacity for a sense of justice and a conception of good), along with the powers of reason (thought and judgment), and has coextensively developed the idea of persons as free and equal.

A parallel evolution can be observed in the law, which first ascribed rights and duties to families, then to tribes, and then to persons, first to men then to women, first to husbands then to wives, first to the healthy then to the ill, first to heterosexuals then to homosexuals (although this latter right is still in process), and so on.

Hence, rights and duties (legal personhood) are a dynamic concept, and the direction of their development cannot be known beforehand. In fact, the current debate on the status of embryos illustrates that we still find ourselves dealing with forms of life whose status as persons has yet to be determined, and that the list of entities eligible for legal personhood might be open-ended.

The content of legal rights and duties depends on the type of legal person these rights and duties apply to. Hence, the rights and duties of humans are different from those of corporations; for example, we humans enjoy some fundamental human rights that corporations do not have.

But there are some features that both natural and artificial persons have in common. These are mainly three: the right to own property and the capacity to sue and be sued [21]. It is these features that bring artificial autonomous agents into play. In fact, the capacity to be sued is why we are discussing these agents and their legal position. If agents were liable, that is, if they could be sued, they would become legal persons, and the task of law would then be to decide whether existing concepts of the legal person (that is, the natural person and the artificial person) can cover artificial agents as well.

So, between natural persons (humans) and artificial ones (corporations), where should we locate artificial agents having the same autonomy as we do?

In what follows I will examine the possibility of considering agents as natural legal persons as against artificial legal persons. But I should point out that this analysis amounts to nothing more than a "thought experiment," as [21] calls it, aiming to "shed light on the debate over the possibility of artificial intelligence and on debates in legal theory about the borderlines of status or personhood."

5 AUTONOMOUS ARTIFICIAL AGENTS AS ARTIFICIAL PERSONS

It is difficult to say which type of personhood is closest to agents because at first glance none of the known legal models of personhood seem to exactly match the personhood of these hypothetical entities of the future. Still, some parallels can be drawn if we take the corporation as an artificial person and use this as a model on which basis to shape a legal perspective through which to conceptualize the hypothetical personhood of artificially autonomous agents.

Four such parallels come to mind. First, just like a corporation, an artificially autonomous agent can be said to *belong* to someone, and this someone can be a natural person, such as a programmer, a software developer, or a user.

Second, just like a corporation is said to live in perpetuity unless it is terminated at the initiative of its shareholders, so the life of an autonomous artificial agent will extend indefinitely unless the agent is put out of existence by its stakeholders (programmer, software developer, user, etc.).

Third, an agent's *liability* can be modelled after that of a corporation, in that its liability for losses or injuries caused to others can be either separated from that of its stakeholders (users and developers) or its stakeholders can be made personally liable, and the parallel here is with a limited and unlimited liability company respectively.

Fourth, there is a parallel to be drawn as concerns the *birth* and *makeup* of the entity in question: just as a corporation comes into existence through a charter (its "birth certificate") providing a broad statement of purpose further defined in the corporation's bylaws, so we can envision the stakeholders of an autonomous artificial agent giving birth to it through a charter and framing its action through bylaws stating what the agent's purpose is, who

its stakeholders are, what its capital structure is, and what its powers are (or what the *extent* of these powers is, which allows for the possibility of *ultra vires* action, offering a framework within which to work out issues of liability).[7]

Still, there is one *but* in considering artificial agents as artificial persons: however we conceive the nature of an artificial person, it will always be *fictitiously* autonomous. A corporation is not *really* autonomous, because its actions are decided by is stakeholders (its shareholders, officers, and directors) and its "will" is always the will of its stakeholders. This is the sense in which artificial persons in the law are considered legal fictions: a corporation is *deemed*, or constructed as, an autonomous person, even though we understand it is not actually autonomous on its own.

Not so in the case of artificial agents: their autonomy is not a fiction; it is *real*, and one of its features is freedom from human control. This is why we cannot strictly ascribe legal personhood to artificially autonomous agents: we cannot assume that these agents express their users' will if we know that agents decide on their own, nor we can assume that someone can control these agents, because these agents act on their own.

Still, although we cannot consider autonomous artificial agents as artificial persons in a strict sense, we will have to concede that the existence of artificial persons in law shows that the law can create new legal forms to welcome novel entries: the development of legal personhood—a status initially ascribed to natural persons and then to artificial ones—shows that the concept of legal personhood can be extended, and in fact that it *was* extended in the effort to meet the need to address technological and industrial developments in the 19th century. Artificial agents may well be the next development of this kind.

In any event, if the autonomy of an artificial agent cannot be properly compared to that of an artificial person—on account of the legal fiction involved in framing the concept of an artificial person—we can still look to other forms of analogy. One idea is that we can think of an artificial entity as a *natural* person, and it is to this idea that I devote the next section.

6 ARTIFICIALLY AUTONOMOUS AGENTS AS NATURAL PERSONS

Natural persons in law are humans, and they enjoy some basic human rights. The question, then, is: Could we, and *should* we, ascribe such rights to artificially autonomous agents?

Basic human rights—"justified, high-priority claims to that minimal level of decent and respectful treatment which we believe is owned to the human being" [22]—include in the first instance the constitutional rights, such as freedom of expression and religion; the right to participate in the political process, as by voting; the right to be secure in one's personal effects; the right to life, liberty, and property; the right to a fair and speedy trial; and the right to be free from "cruel and unusual punishment," to use another well-known phrase; as well as the rights to material subsistence (e.g., the right to health and an opportunity to have gainful employment) and the right to social recognition as an equal member of society.

Undoubtedly, some of the aforementioned rights can only apply to humans, an example being the right to be free from unreasonable search and seizure. Depending on how these rights are conceived, however, they can also be made to apply to nonhuman entities. By way of example, the United States Supreme Court has recently found that corporations and unions can make unlimited campaign contributions (subject to certain restrictions), on the reasoning that a government restriction of such activities would amount to a violation of the First Amendment right to free speech, a ruling that accordingly recognizes that right for corporations and unions.[8] It can thus be argued that humans can and do share some human rights with nonhuman entities—so why can't humans share such rights with artificial agents, too?

There are several arguments why autonomous artificial agents should not be treated as natural persons. In [21] a list of six main reasons is considered suggesting that artificial agents should be precluded from such treatment: agents cannot be treated as humans because they lack (*i*) a soul, (*ii*) intentionality, (*iii*) consciousness, (*iv*) feelings, (*v*) interests, and (*vi*) free will. But the author then proceeds to defeat all these arguments against a "legal anthropomorphization" of autonomous artificial agents: the lack of a soul and of interests (understood as forming the basis for a conception of a good life), he argues, are not valid arguments because we neither agree on what a soul is, nor do we share a common conception of good.[9] The remaining four arguments are defeated by arguing that in each case, "our *experience* should be the arbiter of the dispute: if we had good practical reasons to treat AIs [Artificial Intelligences] as being conscious, having intentions, and possessing feelings, then the argument that the behaviours are not real lacks bite" [21].

In the same vein, [23] argues that sooner or later courts will have to "grapple with the unstated assumption underlying the copyright concepts of authorship and originality, [namely] that 'authors' must be human," while also arguing that "any self-aware robot that speaks English and is able to recognize moral alternatives, and thus make moral choices, should be considered a worthy 'robot person' in our society." Such a robot would have the highest degree of autonomy, such that we would inevitably have to take up the issue of its legal personhood.

These, however, are only the first hurdles an artificial autonomous agent would have to overcome on the path to authentically human behaviour, and there are still many more to come. Just think about the remedies available in dealing with human liability: artificial autonomous agents cannot share liability with humans, because humans can be imprisoned and fined, while artificial agents cannot. True, an agent could conceivably be imprisoned or fined, but such penalties mean different things to humans than they do to agents: imprisonment carries psychological, social, and physical consequences for humans as they do not for agents, while fining imposes on humans a loss that agents cannot suffer, for any money damages would weigh on the agent's owners, not on the agents themselves (unless, that is, we fall back on the analogy of agents

7 There is another kind of analogy that can be struck in thinking about the personhood of an artificial agent: these agents can be analogized not to corporations but to cooperatives, understood as entities created to provide services to its stakeholders, who (where artificial agents are concerned) might be identified as the entire group comprising its users.

8 See *Citizens United v. Federal Election Commission* 558 U.S. (2010), holding that "the First Amendment applies to corporations."

9 A compelling statement of this argument is offered in [6], noting that what we share is not a single broadly accepted moral conception but a sparse collection of generally accepted moral norms.

as artificial persons). For these reasons the natural-person analogy does not quite help us solve the problem of how artificial agents should be treated from a legal perspective.

7 CONCLUSIONS AND FUTURE WORK

So what conclusions can we draw from the foregoing discussion? One thing is clear, that even when CS will advance to the point where it enables us to build an artificial agent that is fully autonomous in all senses of this term—Kantian and computational—the problems to be solved will not come to an end but will on the contrary multiply. It may very well be that we can work out these issues as we go along, but I believe that it is nevertheless important to think about the implications ahead of time. I believe that the more we discuss them, the greater the likelihood that we will have ideas, insights, and solutions we can put to use so as to be ready in time. As [6] argues, the law has an advantage over other disciplines in working toward practical solutions to the legal and moral responsibility of artificial agents, precisely because the law is accustomed to dealing with such practical problems, and so we should persist in our effort, not dismissing any avenue of research as too far-flung.

In the meantime, we still have to ask: How might the law proceed in treating artificially autonomous agents if it cannot apply to them either of the two forms of personhood, the natural or the artificial? One suggestion I would have is that of hybrid personhood: a quasi-legal person that would be recognized as having a menu of rights and duties selected from those we currently ascribe to both natural and artificial persons, the idea being that we need not commit to any one analogy in working out the question of an artificial agent's autonomy and liability. Unfortunately, there are quite a few sizable obstacles that will need to be overcome in pursuing such an approach: to begin with, we would have to come up with an appropriate list of rights and duties, then we would have to decide which of these rights and duties apply—depending on the different areas of activity and the different types of agents involved—and finally we would have to work out agreed procedures for deciding how these rights and duties are to be applied and who will be empowered to make such decisions. But, as I suggested a moment ago, this is very much a work in progress: the beauty of it is that, although we may not have all the solutions ready at hand, it is probably not advisable to attempt a comprehensive theory before we even know what an autonomous artificial entity will exactly look like, because we can probably develop better insights as we go along, provided we do not become complacent and set the problem aside entirely, thinking that we can solve it when it becomes real.

REFERENCES

[1] I. Kant. *Critique of Pure Reason,* Dover Publications, USA, ([1781] 2004).

[2] J. B. Schneewind. *The Invention of Autonomy*, Cambridge University Press, UK (1998).

[3] L. Floridi and J. W. Sanders. On the Morality of Artificial Agents, *Mind and Machine,* 14: 349–379 (2004).

[4] C. Allen, G. Varner and J. Zinser. Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence,* 12:251–261 (2000).

[5] K. G. Coleman. Android Arete: Toward a Virtue Ethic for Computational Agents. *Ethics and Information Technology,* 3:247–265 (2001).

[6] P. M. Asaro. What Should We Want From a Robot Ethic? International Review of Information Ethics, 12(6): 916 (2006).

[7] G. Dworkin. *The Theory and Practice of Autonomy*, Cambridge University Press, USA, (1988).

[8] W. Wallach and C. Allen. Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, USA (2009).

[9] T. Smithers. Autonomy in Robots and Other Agents. *Brain and Cognition,* 34:88–106 (1997).

[10] A. A. Covrigarand and R. K. Lindsay. Deterministic Autonomous Systems. *Artificial Intelligence Magazine,* 12 (3): 110–117 (1991).

[11] Royal Academy of Engineering. *Autonomous Systems: Social, Legal and Ethical Issues.* Royal Academy of Engineering, UK, (2009).

[12] R. Siegwart and I. R. Nourbakhsh. *Introduction to Autonomous Mobile Robots.* MIT Press, USA, (2004).

[13] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, USA, (2003).

[14] G. A. Bekey. On Autonomous Robots. *The Knowledge Engineering Review,* 13(2):143–146 (1998).

[15] A. A. Hopgood. *Intelligent Systems for Engineers and Scientists,* CRC Press, USA (2001).

[16] K. P. Sycara. The Many Faces of Agents. *AI Magazine,* 19(2):11–12 (1998).

[17] J. Odell. *Introduction to Agents*. http://www.objs.com/agent/agents_omg.pdf. (2000).

[18] D. J. Calverley. Imagining a Non-Biological Machine as a Legal Person. Artificial Intelligence & Society, 22(4): 523–537 (2008).

[19] R. W. Picard. *Affective Computing*, MIT Press, USA, (1997).

[20] C. Darwin. *The Descent of Man*. Penguin Classics, UK ([1871] 2004).

[21] L. B. Solum. Legal Personhood for Artificial Intelligences. *North Carolina Law Review*, 70: 1231–1287 (1992).

[22] B. Orend. *Human Rights: Concept and Context*. Broadview Press, Canada, (2002).

[23] A. R. Freitas Jr. The Legal Rights of Robots. *Student Lawyer*, 13: 54–57 (1985).

[24] J. Raz. The Morality of Freedom, Clarendon Press, UK (1988).

[25] J. Rawls. Political Liberalism, Columbia University Press, USA, (1996).

[26] R. Cohon. Hume's Moral Philosophy. *Stanford Encyclopedia of Philosophy* (2010).

Socialness in man-machine-interaction and the structure of thought

Bernhard Will[1] **and Gerhard Chr. Bukow**[12]

Abstract. We propose that socialness in man-machine-interaction is reached only in a cognitively informed way and bring in different results from philosophy and psychology to handle the structure of human belief in social interaction adequately.

1 MASTERING THE TURING TEST

The Turing test is expected to be a measure judged by humans for a machine's intelligence, socialness or humanness of cognition in a *dialogical* man-machine-interaction scenario. Many issues have been treated regarding the "strange" foundations of Turing test-interactions: is it really social without embodied contact or shared aims? Or, the dependence from the subjectivity of interpretation of humanness. However, the Turing test promotes full-flagged functionalism regarding the material realization of the machine – and so the machine's judges will refer to issues like expert knowledge, daily experiences, or content coherence, and the existence of own points of view. For many people (and especially lay people), it may be straight forward to think about essential "human universals" that should generate dialogical behavior. But is this the only approach?

From an engineering point of view, it might be quite clear that another approach could generate the "most human" dialogue sequence: given enough interactions (at best, infinitely many ones), using a statistical method like Markov-chains will give you the most probable "most human" messages depending on the history of interaction. This approach is essentially poor of theory and is comparable to a situation well known in the astronomy of the middle ages: given a very high or probably infinite number of spheres, there could be a best model describing the orbits of all planets in the universe. If there are problems with the predictions generated by the model, one just has to add another sphere influencing the other spheres. But like the Markov-chains-approach that could be used for the Turing test without having ever used a theory about human essentials, you could do this without having ever captured the theory about the essentials of the physics of universe. You would only look for non-explanatorily surfaces that do not capture human belief.

This story about surface and generating structure tells us two issues: 1) dialogues are usually seen in a contentful manner by Turing test-judges. But they might concentrate on the surface structure of contents (like the spheres) without caring for the essential structures of thought that generate content; 2) work about the Turing test – or any other man-machine-interaction – could be done without any humanly informed way. We propose that a "social turn" should require a cognitively informed, predictive *and* explanatory way that handles human cognition and its constraints on dialogical interaction.

1Institute of Philosophy, University of Magdeburg, Germany

2Institute of Psychology, University of Giessen, Germany

2 DIALOG-BASED ISSUES IN TURING TESTS

The surface-structure of a Turing test is based on the sequences of dialog messages. These messages may represent beliefs held by agents engaging in a dialogue. Next to our last question, whether we should realize this engagement in an informed way, it is a serious issue, whether we should concentrate on the content (i.e. surface) of dialogical sequences or on its underlying belief structures. Both options are integrated in a sequential model of surface structures, but the moves and consequences within this model are modeled very differently. Let us first look at the content-based option and then consider the structural option in the next chapter.

Sequences of contents are driven purely semantically, by world knowledge or other contentual strategies. A "good" dialogue should have content typed "human". Belief contents can be inspected with means of coherence: is the dialogue story coherent? Are all the parts of the story explanatorily relevant for each other? Are measures of information distribution and interchange rate in normal ranges of communication?

But the focus on sequences of contents alone has serious deficits:

- It seems hopeless to wait for a purely "semantically driven" theory of content that guides you just from content to content while only respecting content. This theory would also imply a solution for the problem of the relation between syntax and semantics, which is really hard.
- The focus on content is typically expressed by the hypothesis that agents work in a propositional format. However, every proposition is believed in a representational format.
- The focus on "linear sequences" of contents may neither respect the "holistic" structure of thought nor the properties that guide acquiring, abandoning or revising belief. It may seem that content alone would be relevant for the "next" belief, but it is commonly known within the cognitive community that structures of thought are also relevant.

So, let us hold that the content-driven *sequence-model* in the Turing test-debate has some deficits concerning the capturing of belief-based human cognition. We propose that attention to the structural approach can support some fixes of these deficits and want to motivate this by considering the reverse Turing test.

3 THE REVERSE TURING TEST

The reverse Turing test lets us think about how one should generate the sequence of belief from the point of view of the machine that should test for humanness. It is clear that we cannot ad hoc assume that machines can focus on the contents of belief or on intentions linked to representations, because one would already assume the intentionality and "humanness" of the

machine – which is circular with respect to our problem. However, a machine only driven by probabilistic methods, would be problematic, too: the machine would already be expected to be without any content-component. This would undermine our models of humanness.

So, how could we figure out the desired humanness and ability for social interaction with humans to be implemented in the Turing test? It is worthy to have a look to philosophy of science where three general types of strategies are known: 1) list strategy 2) universals strategy 3) structures strategy.

1. The list strategy seeks for a list of desired features of humanness or socialness. However, this approach is most problematic, since we need another list of criteria to legitimate each list, which is circular.
2. The universals strategy seeks for what is common among all humans (i.e. universals). But "content-universals" have a difficult standing with respect to actual human psychology. The Maslow pyramid of motivations does show this: though the pyramid is intuitively appealing and does suggest several motivations (i.e. contents), empirically, it has been shown to be quite worthless. Instead of content theories, theories of processes are successful and adequate to describe human behavior, and e.g. their selection of specific contents.
3. The structures strategy focuses on the structures generating the surfaces of content. This is our second mentioned option of the last chapter. In the view of the deficits of the other options, we will consider two different structures underlying the Turing test: 1) the man-machine-interaction model 2) examples for the structure of human thought and how it might play a role in dialogical interaction. This also is the adequate strategy for the machine that cannot know intrinsically any human universals or cannot legitimate a special list.

The Reverse Turing test-perspective now shall be combined with the structural strategy to investigate, how a machine could "realize" human moves of thought. To do this, we consider two more preliminary points regarding first the planning framework and second the right type of explanation we should expect.

4 "JOINT", "SHARED" AND EXPLANATIONS

Dialogical models of man-machine-interaction usually take a planning approach and add individual agents that collaborate in terms of planning and acting with respect to shared goals or joint awareness of the environment (e.g. being aware of each other). In this view, coherent verbal dialogues are just special cases of mutually planned actions. "Joint" or "shared" is regular "planning-babbel" that can be viewed from at least two opposite positions with respect to the notions of global or local standards of intentional planning.

Some researchers, like Bratman [1], take individual plans to be globally "meshed" such that e.g. a dialogue can be reduced to intentions, actions and their organization. Common activities are reduced to intentions and meshing delivers necessary and sufficient conditions for social interaction (conditions to speak about social interactions at all) such that socialness depends on plan meshing. Some other researchers claim that a shared activity with a shared intention is irreducible to the individual intentions of the participants; however, we do not follow this irreducibility here. In these cases, the social interaction has a *global* nature: individuals share intentions and plans from a global point of view that also regulates the sub-global points of view. This claim about the necessity of globalist plan- meshing is far too strong and unsuitable for the description of man-machine-interaction, and it is too global to be achievable for a machine without life-long history of interaction.

Instead, we argue that neither shared intentionality nor plan - meshing is required for a successful interaction.

Agreeing with Hollnagel [2] on joint cognitive systems and control, and Suchman's [3] view on situated action, successful interaction requires the machine's ability to recognize and support the intentions of the user at a local and situational level. To be able to fullfil these tasks, the machine has to be cognitively informed to cope with the user's mistakes and intentions – to investigate in these abilities with respect to machines, we have to consider the Reverse Turing test.

Actual research does not take into account interaction from a Reverse Turing test perspective: mostly, psychology discusses human-human-interaction and computer science is concerned with human-machine-interaction from the view point of a human being. What type of explanation could be useful and adequate for this type of perspective? In the framework of planning resp. planned dialogical interaction, we can already dismiss statistical explanations. No statistical explanation given by a machine without taking the human "structural" perspective into account delivers an acceptable explanation for humans – it does not provide reasons. We should also take care with too vaguely formulated types of mechanisms, e.g. "neuro-cognitive mechanisms" (see e.g. Sebanz et al. [4]). These mechanisms implicate something like a "nomological bridge" between cognition and neurological realization that is excluded in principle within functionalism promoted by the Turing test.

The right level for our problem is the cognitive level (e.g. promoted by classic cognitivism in the sense Jerry Fodor has promoted it) that is committed to functionalism and properties of cognitive systems (e.g. representationalism, systematicity, etc.). Cognitivism also suggests a specific type of explanation that is based on the functional role of a cognitive entity. Such an entity has its role in a network of entities and this network is configured in specific ways to fulfill specific tasks. Let us apply this type of explanation now in our consideration of human structures of thought.

5 STRUCTURES OF THOUGHT

We now want to consider three examples that show specifically human ways of structures of thought that should be taken into account in the attempt to generate socialness in man-machine-interaction: 1) belief systems and their changes; 2) special representational formats of beliefs and their specific ways of changes, e.g. mental models and their variation with respect to preference and epistemic equivalence; 3) epistemic accessibility and the explicit/implicit-distinction.

All of these aspects normally depend on some very specific pictures of the maximally rational agent. We already had considered a rational agent in the case of global notions of planning and interaction. However, we have good reasons to take into account realistic models of agents when considering the structure of human thought from a Reverse Turing test-perspective. If we aim at a positive explanation for successful social interaction, the most acceptable explanation will certainly not be the mention of deficits of actual humans to a fundamentally different and idealized cognitive agent. There are infinitely fundamentally different models we could take into account – but why should any of these ones matter if we judge the humanness of a machine or the machine should judge the humanness of a potential human? As far as we can imagine, the only useful way would be a "metric" to measure the distance between actual human and all idealized rational agent models. However, as far as we know, no such metric for rational agents has been suggested. So, let us consider three examples where humans diverge from ideal agents and ideal rationality.

1. Belief systems and their change

The change of belief systems is generally analogical to theory change – a well known topic in philosophy of science. Which beliefs (or laws or entities in case of theory) should be adopted, abandoned or acquired in the front of new information (or confirmation/disconfirmation etc.)? The change of belief systems is a typical feature of agents that are not omniscient with respect to the world and logic. Neither do they know everything, nor do they believe in any consequence of their already established beliefs, nor do they have unlimited computational capacities.

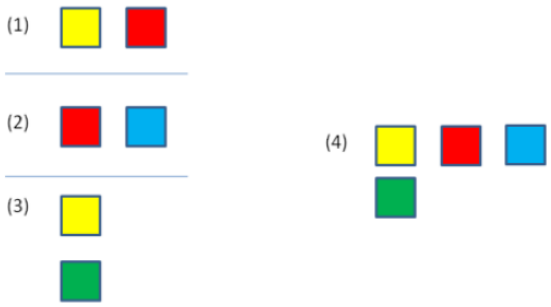
Depending on one's epistemological position, there are some different frameworks one can choose for the norms and descriptions of rational change of belief. We do not need to go into detail with respect to any of these approaches and their differences. However, their common feature is that typically the change of belief is not uniquely determined. There is actually no theoretical framework that provides norms and descriptions for this determination as well as for iterated change (in case of revision). Furthermore, with respect to actual humans, change depends on several features, sensitive to context, semantics, and syntax, as well as epistemic/doxastic features like preference, equivalence, and representational format.

In case of Turing test or Reverse Turing test, we propose a strong link between the way belief systems do change and the judgment of humanness of the produced sequence. A good practical example is delivered by the theory of mental models and its experimental apparatus that we want to consider now.

2. Mental models, preferences and epistemic equivalency

Belief revision typically assumes that beliefs are just the propositions expressed and thought in language-like sentences. Our cognitive abilities – e.g. the ability to infer – are just thought to be possible because of inferential relations between propositions. But the kingdom of mental representations and their properties is much larger than sentences expressing propositions alone. We have some reasons to accept this: if one takes seriously the insight that propositions are always believed in a representational format and that the representational formats of working memory and long term memory do differ. A consequence of these two insights is that change of belief can differ with respect to format issues. Current research (e.g. Jahn et al. [5]) investigates the construction and revision (there called variation) of mental models in the realm of cognitive psychologist Johnson-Laird. These models are built upon propositions that describe spatial scenes, but can be used for every scene that integrates relational information. After building up the model, cognitive processes work on the mental model. Additional scanning procedures then scan the model to generate new propositions describing the scene in the model.

The following pictorial example taken up from the BELIEF SPACE project led by Markus Knauff at University of Giessen gives an impression of how to construct from premises (1), (2) and (3) a mental model (4). The premises give relations between things located at several places such that we are in the area of spatial reasoning. However, you may use other items with different complexity or semantics as well as other relations, too.



In the light of new evidence or information inconsistent with the model (here: (a)), however, the model has to be changed.

There are two possibilities to revise the model if (a) is presented such that it is new information relevant for the model: (a) and (b).

Now, cognitive research shows that – given several logically equivalent ways to construct and to revise a model – some ways (i.e. models) are preferred. These preferences are constant within individuals and within groups and show that there are cognitively significant aspects that are not captured by the logical description alone. This does not mean that humans prefer in an "illogically" way – but that epistemic equivalence may have to do with other equivalence-relations than classic logical ones.

Preference and equivalence do play a major role in the generation of beliefs and an agent's ability to track them.

However, as the next point shows, not all beliefs are "assessable" for the human agent. The ability to follow the generation of preferred models is essential for socialness. Just imagine, how "social" a group of human agents would be if they do not *see* the model of other group members!

3. Epistemic accessibility and coherence

Epistemic logic assumes that the cognitive agent is able to overview his own beliefs in two important senses: (1) the agent is fully aware of all beliefs (2) the agent knows and believes every consequence of his set of beliefs. Both assumptions are critical, because either we take them seriously and neglect certain aspects of real agents, or we discard these assumptions and have to discard standard ways to model agents with the help of epistemic logic, too. It is an open question how to model realistic belief structures of actual humans without using something like an awareness-function that "signs" every belief we are aware of. Of course, the problem is how such an awareness-function would be formulated and if it is psychologically adequate. There are some alternatives to classic logics without such functions, but these do have their own problems (of course). But again we just need to consider the principle problem.

Let us consider coreferential situations of names, that is a type of situation where one may have some implicit knowledge from a formal point of view, but cannot access it. Julia may know that Cicero is a great orator, and she may also know that Tully is a great orator. However, the reference may be opaque such that Julia may not know that *Cicero = Tully* holds. But, in a certain sense (called direct reference), Julia may know implicitly that Cicero = Tully because if her beliefs refer to truthful circumstances, she refers to Cicero in cases of believing something about Cicero and in cases of believing something about Tully. But this is not assessable and for this reason she may also not believe the consequences of these beliefs. If Julia gets the information that Cicero = Tully (e.g. by analogical inference or by seeing Cicero when she expects Tully to talk), her implicit knowledge may be assessable for her.

How should we model and understand this shift in epistemic accessibility? From the revision point of view, it may seem that Julia "revises" her belief set by a new fact "Cicero = Tully". But this cannot be the case if belief revision assumes that Julia's beliefs are referring to the world. Julia knows – in a way – this information already and it is not a real change. And, the cognitive dimension of the *expansion* of accessibility from implicitness to explicitness may not be described adequately as a *revision*. Julia does not have explicit false beliefs about Cicero and Tully. *Doxastic logic* [6] may be the most promising alternative in terms of logics, because it does not require epistemic closure and has a notion of equivalence. It does not imply awareness functions. However, we cannot detect coherence directly and modeling with doxastic logic has its own difficulties, if we want to respect cognitive information that is not relevant for doxastic actions (e.g. doxastic logic is format-neutral).

An alternative suggestion how to "detect" a change in epistemic accessibility may be a change of *coherence* in the time line of the modeled agent. It seems obvious that both the coherence of elements and the coherence of the whole belief network are higher if Julia believes that "Cicero = Tully". Analogy of explanations (Cicero does this, Tully does this, Cicero was here, Tully was here …) is a coherence relation such that there are not two isolated "blocks" without relations named like "Cicero" and "Tully" do exist. More epistemic accessibility means more coherence with respect to the ordinary belief set handled in epistemic approaches, and it can be vindicated by e.g. following actions depending on coherence. However, this coherence is not only a content feature (like in casual dialog models) – it is a feature of the accessibility-structure of human thought. This can easily be modeled e.g. with EchoJAVA without implementing directly the implicit hypothesis "Cicero = Tully" (H3) or "Cicero != Tully".

If machines shall act in the realm of socialness – that is being able to take part in social interaction and understand social situations – both knowing how humans could have implicit beliefs and how these beliefs may come explicit and causally efficient are important to understand behavior in social circumstances.

Table 1. Simple coherence properties of epistemic accessibility in JavaEcho http://cogsci.uwaterloo.ca/JavaECHO/jecho.html, see e.g. Thagard (2000).

Coherence without H3: 0,037192	Coherence with H3: 0,049084

// H1 - Cicero is a great orator
// H2 - Tully is a great orator
// H3 - Cicero equals Tully
// E1 - see Tully
// E2 - see Cicero
 	
contradict(H1,E1)
contradict(H2,E2)
explain((H1),E2)
explain((H2),E1)
	// H1 - Cicero is a great orator
	// H2 - Tully is a great orator
	// H3 - Cicero equals Tully
	// E1 - see Tully
	// E2 - see Cicero
	
	contradict(H1,E1)
	contradict(H2,E2)
	explain((H1),E2)
	explain((H2),E1)
	
	explain((H3),E1)
	explain((H3),E2)
	analogous((H3,H1),(H3,H2))
	analogous((H1,E1),(H2,E2))

6 CONCLUSIONS

Let us make three conclusions based on our treatment of dialogical situations in man-machine-interaction and the background problem of the (reverse) Turing test:

1. Socialness is not just a feature of content or sequences of content. It is also a feature beared by the structure that generate beliefs having such contents and guide belief systems in case of change, accessibility, preference, equivalence, representational format, and other features. These features are proposed to be necessary conditions for social cognitive agents that deserve their labels, at least in social interaction with humans.

2. We should consider not only the Turing test-situation, but also other situations of planned social interaction from different perspectives, e.g. the Reverse Turing test-situation. These perspectives force us not just to take care for content-based research and circular assumptions of content-understanding machines, but also for structural aspects and content-understanding capabilities from scratch up.

3. However, we should be aware of the right level and type of explanation: cognitive explanations do provide a way to inform us and our machines about the typical defaults of humanness and socialness. These cognitive aspects cannot be reduced to statistics or brute force in the long run.

REFERENCES

[1] Bratman, M. (1999). *Intention, Plans, and Practical Reason*. Cambridge University Press.

[2] Hollnagel, E. (1983). What we do not know about man-machine systems. *International Journal of Man-Machine Studies, 18, 2, 135-143.*

[3] Suchman, L. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press.

[4] Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Science, 10, 2, 70-76.*

[5] Jahn, G., Knauff, M., & Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory & Cognition, 35, 2075-2087.*

[6] Belnap, N., Perloff, M., & Xu, M. (2001). *Facing the future*. Oxford: Oxford University Press.

[7] Thagard, P. (2000). *Coherence in Thought and Action*. MIT Press.

Virtual Sociality or Social Virtuality in Digital Games? Encountering a Paradigm Shift of Action and Actor Models

Diego Compagna[1]

Abstract. In this paper, I argue that digital games are a best case scenario for new forms of action and especially for new actor models. Social computing is not just about humans bringing the social world into virtuality or finding some sort of social terms in the virtual environments, but constitutes a way that, as social actors, humans are reshaped by the new forms of social realities (even if we find them within virtuality). In Mead's definition of action and actor model, the meaning of a symbol (and, to that effect, the meaning of one's own thoughts and view, and finally one's 'sense of self`) depends on the reaction of the other (alter). The meaning of a symbol constitutes ex post according to alter's reaction to it. In these terms, 'knowing' something means to anticipate alter's (most probable) reaction/understanding. In the end, this means that a clear distinction between the player and his or her avatar cannot be presumed. As a cybernetic feedback loop, they create a oneness or an integrated interface: The avatar and the player (at least as long as he or she is playing) are social actors within the game-play space, even if the player is physically located outside the virtual environment of the game-play space - in almost the same manner as Luhmann (relying on Mead) claimed that the actor's mind is outside the environment of the interaction.

1 INTRODUCTION

The relationship between the player of a digital game and his or her avatar (or model)[2] is intriguing: Who or what is actually acting and where does the action take place? A first step towards clarifying this peculiar situation is to distinguish between two different areas of action. The first area involves the human player using fine motor skills and usually takes place within a one meter radius at most. The second is the player's avatar's area of action, which can cover great distances depending on the type of game [8]. Especially in comparison with the player's range of actions and motions, the avatar is usually capable of performing a much wider variety of actions, usually characterized by a very high degree of freedom [15, 12].

The literature describes three areas or dimensions related to this topic: 1.) physical space: this is the human player's space (α); 2.) game-play space: this is the virtual environment where avatars act (β); 3.) social-symbolic space: this is the space where social interactions take place and social meaning/symbols are used or emerge (γ). The crucial point I would like to emphasize is that some scholars dealing with digital games locate the area where symbolically mediated interaction takes place within the game-play space, i.e. within the avatar's virtual surroundings and very far away from the 'human's' location [5, 9, 15]. This clearly gives us cause to assume a new form of sociality within the virtual (i.e. virtual sociality). Then again, maybe characterizing this phenomenon as a form of virtual sociality is misleading - if an entity's interaction can be described as a symbolically mediated one, the effects are the real construction of social worlds. For these purposes, is it still appropriate to call it 'virtual' by any means?

2 ACTING WITH/ IN DIGITAL GAMES

Britta Neitzel examines the relationship between player and avatar by positing a distinction between a "point of action" (PoA) and a "point of view" (PoV) [11, 15]. First, Neitzel assumes that the connection between player and avatar can be characterized by very tightly wound feedback loops or cybernetic models [9]. Second, she asserts a strict division between player (α) and avatar (β), due to the fact that the player's perspective (PoV) is outside the game-play space (PoA) [9, 11]. Although the player acts within the game-play space (PoA, β), he or she remains outside of this area and stays planted in the player's (i.e. the human's) location (PoV, α) [10]. Due to the fact that the player is constantly observing (PoV) his or her avatar and its actions within the game-play space (PoA).

Although the player acts in the game-play space through his or her avatar, he or she is constantly aware that the avatar is merely a representative performing actions in a purely virtual environment (qualitatively different and strictly separate from the player's reality) [10]. Neitzel attributes this differentiation to the human player's observation of his or her avatar, even though she characterizes the game-play space (β) as the area where symbolically mediated interactions take place (γ) [11, 9]. Neitzel refers to George Herbert Mead's action theory in describing the game-play space as the area where symbolically mediated interactions take place [9]. Under these circumstances it becomes quite peculiar to argue that the avatar (or more precisely, the human player observing his or her avatar) is grounds for differentiating between the two areas. Even if Neitzel states that the relationship between player and avatar could be described as a pair of entities strongly connected by cybernetic feedback loops, the two entities remain strictly separated.

I would like to stress that characterizing the game-play space (β) with Mead's concept of symbolically mediated interaction (γ) could or should lead to a completely different conclusion

[1] University of Duisburg-Essen, Faculty of Social Sciences, Germany Email: diego.compagna@uni-duisburg-essen.de

[2] In this paper the expression "avatar" will be used to refer to the (virtual) game character, inde-pendent of genre. In shooter games this is often referred to as a "model" and in role playing games an "avatar."

regarding the relationship between the player (α) and his or her avatar (β = γ). In Mead's definition of action and actor model, the meaning of a symbol (and, to that effect, the meaning of one's own thoughts and view, and finally one's 'sense of self') depends on the reaction of the other (alter). The meaning of a symbol constitutes ex post according to alter's reaction to it. In these terms, 'knowing' something means to anticipate alter's (most probable) reaction/understanding. Mead emphasizes the so-called 'vocal gesture' because humans have the physiological ability to hear the 'spoken symbol' (e.g. word) in the same way and at the same time as alter [7]. From a biological and physiological point of view, language played a useful role in social evolution as a tool for successful interactions. Applying this concept to the previously mentioned situation, one can easily see the strong parallel: The player is able to observe his or her own actions at the same time as alter (the player of another avatar) is seeing them. The player anticipates his or her action (mediated by his or her avatar) in a very similar way to Mead's description of the vocal gesture.

3 SYMBOLICALLY MEDIATED INTERACTION AND THE RECONCILIATION OF 'POV' AND 'POA'

The player (ego) is able to anticipate the view of his or her teammate (alter) not just because he or she is able to hear what he or she is saying to alter, but also because ego can actually see his or her own avatar acting in just the same way alter sees it. The accentuated weight of the vocal gesture can be easily transferred to bodily related gestures. As a matter of course the importance of the vocal gesture plays a fundamental role in Mead's theory on a very basic level. It describes the connection between the ontogenesis and the phylogenesis of the "social" that can be traced back to the effects of symbolically mediated interaction and the (intersubjective) social reality it constructs [7]. Nevertheless, by transferring Mead's general and abstract concepts that link action, sociality, and identity to the concrete phenomenon of digital games, one can easily conclude that making a distinction between the PoV from the PoA leads to the exact opposite of Neitzel's deduction.

In the end, Mead's action theory is also the core model for Niklas Luhmann's micro-level theory of social systems (interaction system) and could be used to explain how consciousness is linked to the social world (both, of course, as systems): The ego's psychological system (self-awareness, consciousness) is constantly observing the interaction between alter and ego, but it remains in the environment of the interaction/social system [6, 4]. The mere proposition of ego observing the interaction in which he or she is involved does not mean that a clear distinction or some sort of 'border' keeping the player apart from his or her avatar can be presumed. Quite the contrary, especially if the situation is described using Mead's theory. Mead's complex social explication of action and the way social actors' identity and self-awareness is bound with symbolically mediated interactions is deeply misunderstood by Neitzel. Her argumentation is based on the differentiation between the PoV and the PoA, although according to Mead or Luhmann there is no PoV that is not decidedly intertwined with the location where the action is taking place: The PoA (β) is the only area where social meaning can possibly emerge (γ), which, in turn, gives rise to self-awareness and -consciousness (α), which made a PoV possible.

Some of the cybertext approaches compared to Neitzel's view are much closer to my view: The avatar is an essential part of the feedback loop that constitutes the player as an actor [3]. Metaphorically speaking, one can say that the avatar becomes a prosthesis of the player [1]. Unlike Neitzel's view (which could be seen as a showcase for monolithic actor models), the game-play area cannot be separated from the actor, who in turn is constituted by his or her actions performed by his or her avatar. In the end, this means that a clear distinction between the player and his or her avatar cannot be presumed. As a cybernetic feedback loop, they create a oneness or an integrated interface [2]: The avatar and the player (at least as long as he or she is playing) are social actors within the game-play space, even if the player (and this certainly applies to the PoV as well) is physically located outside the virtual environment of the game-play space - in almost the same manner as the actor's mind is outside the environment of the interaction. Finally, the differentiation between PoV and PoA is completely irrelevant in terms of describing or achieving better understanding of the question at stake. Of course the situation can only be described this way if the player is able to experience an immersion in the flow of game-play. To do so he or she must be able to control his or her avatar in a similar way how he or she have learned how to move his or her body [14, 13, 9].

4 CONCLUSION

In this paper, I argue that digital games are a best case scenario for new forms of action and especially for new actor models. Social computing is not just about humans bringing the social world into virtuality or finding some sort of social terms in the virtual environments, but constitutes a way that, as social actors, humans are reshaped by the new forms of social realities (even if we find them within virtuality).

REFERENCES

[1] K. Bartels. Vom Elephant Land bis Second Life. Eine Archäologie des Computerspiels als Raumprothese. In: Hamburger Hefte zur Medienkultur 5, pp. 82-100 (2007).

[2] J. Baudrillard. Videowelt und fraktales Subjekt. In: Ars Electronica (Hg.): Philosophien der neuen Technologie. (1. Aufl.) Berlin: Merve-Verl. pp. 113-131 (1989).

[3] T. Friedman. Making sense of software. Computer games and interactive textuality. In: Jones, Steven G. (Hg.): CyberSociety. Computer-mediated communication and community. (1. Aufl.) Thousand Oaks [u.a.]: Sage Publications. pp. 73-89 (1995).

[4] A. Hahn. Der Mensch in der deutschen Systemtheorie. In: Bröckling, Ulrich / Paul, Axel T. / Kaufmann, Stefan (Hg.): Vernunft - Entwicklung - Leben. Schlüsselbegriffe der Moderne. Festschrift für Wolfgang Eßbach. (1. Aufl.) München: Fink. pp. 279-290 (2004).

[5] A. Kerr. The business and culture of digital games. Gamework/Gameplay. (1. Aufl.) London [u.a.]: SAGE (2006).

[6] N. Luhmann. Soziale Systeme. Grundriß einer allgemeinen Theorie. (6. Aufl.) [Original: (1984)] Frankfurt a.M.: Suhrkamp (1996).

[7] G. H. Mead. Geist, Identität und Gesellschaft. Aus der Sicht des Sozialbehaviorismus. (13. Aufl.) [Original: (1934)] Frankfurt a.M.: Suhrkamp (2002).

[8] B. Neitzel. Die Frage nach Gott. Oder warum spielen wir eigentlich so gerne Computerspiele. In: Ästhetik und Kommunikation 115, pp. 61-67 (2001).

[9] B. Neitzel. Wer bin ich?. Thesen zur Avatar-Spieler Bindung. In: Neitzel, Britta / Bopp, Matthias / Nohr, Rolf F. (Hg.): "See? I'm real.". Multidisziplinäre Zugänge zum Computerspiel am Beispiel von 'Silent Hill'. (1. Aufl.) Münster [u.a.]: LIT. pp. 193-212 (2004a).
[10] B. Neitzel. Gespielte Geschichten. Struktur- und prozessanalytische Untersuchungen der Narrativität von Videospielen. [Original: (2000)] Weimar: Dissertation, Bauhaus-Universität Weimar, Fakultät Medien (2004b).
[11] B. Neitzel. Point of View und Point of Action. Eine Perspektive auf die Perspektive in Computerspielen. In: Hamburger Hefte zur Medienkultur 5, pp. 8-28 (2007).
[12] R. F. Nohr. Raumfetischismus. Topographien des Spiels. In: Hamburger Hefte zur Medienkultur 5, pp. 61-81 (2007).
[13] C. Pias. Computer-Spiel-Welten. (1. Aufl.) München: Sequenzia (2002).
[14] J. Sleegers. Und das soll Spaß machen?. Faszinationskraft. In: Kaminski, Winfred / Witting, Tanja (Hg.): Digitale Spielräume. Basiswissen Computer- und Videospiele. (1. Aufl.) München: Kopaed. pp. 17-20 (2007).
[15] J.-N. Thon. Unendliche Weiten?. Schauplätze, fiktionale Plätze und soziale Räume heutiger Computerspiele. In: Hamburger Hefte zur Medienkultur 5, pp. 29-60 (2007).

A Multi-Dimensional Agency Concept for Social Computing Systems

Sabine Thürmel[1]

Abstract. In order to understand agency and interagency in virtual and hybrid constellations the state of the art in attributing collective and distributed agency in socio-technical systems is outlined. A concept of multi-dimensional, gradual agency is introduced and its applicability to social computing systems is demonstrated.

1 POTENTIALITY AND ACTUALITY OF SOCIAL COMPUTING SYSTEMS

Computer simulations let us explore the dynamic behaviour of complex systems. Today they are not only used in natural sciences and computational engineering but also in computational sociology. Social computing systems focus on the simulation of complex interactions and relationships of individual human and/or nonhuman agents. If the simulations are based on scientific abstractions of real-world problem spaces they enable us to gain new insights. "Crowd simulation" systems are useful if evacuation plans have to be developed. Demonstrators for the coordination of emergency response services in disaster management systems, based on electronic market mechanisms, have been built [1].

Computer-based simulations provide a link between theory and experiment. Social simulation systems are similar to numerical simulations but use different conceptual and software models. Numerical methods based on non-linear equation systems support the simulation of quantitative aspects of complex, discrete systems [2]. In contrast, multi-agent systems (MAS) [3] permit to model collective behaviour based on the local perspectives of individuals, their high level cognitive processes and their interaction with the environment. Both approaches may complement each other. They can even be integrated to simulate both numerical, quantitative and qualitative, logical aspects e.g. within one expressive temporal specification language [4]. Agent-based models (ABMs) may be better suited than conventional economic models to model the "herding" among investors. Early-warning systems for the next financial crisis could be built based on ABMs [5]. The Agile project (Advanced Governance of Information services through Legal Engineering) is even searching for a Ph.D candidate to develop new policies in tax evasion scenarios based on ABMs [6]. The novel technical options of "social computing" do not only offer to explain social behaviour but they may also suggest ways how to change it.

Simulations owe their attractiveness to the elaborate rhetoric of the virtual [7]: "It is a question of representing a future and hypothetical situation as if it were given neglecting the temporal and factual dimensions separating us from it – i.e. to represent it as actual" [8, p.4]. Social computing systems are virtual systems modeled e.g. by MAS and realized by the corresponding dynamic computer-mediated environments.

Virtuality in technologically induced contexts is even better explained if Hubig's two-tiered presentation of technology in general as a medium is adopted. He distinguishes between the "potential sphere of the realization of potential ends" and the "actual sphere of realizing possible ends" [9, p. 256]. Applied to social computing systems it can be stated that their specification corresponds to the "potential sphere of the realization of potential ends" and any run-time instantiation to a corresponding actual sphere. In other words: Due to their nature as computational artifacts the potential of social computing systems becomes actual in a concrete instantiation. Their inherent potentiality is actualised during runtime. „A technical system constitutes a potentiality which only becomes a reality if and when the system is identified as relevant for agency and is embedded into concrete contexts of action" [9, p.3].

Since purely computational artifacts are intangible, i.e. existing in time but not in space, the situation becomes even more challenging: one and the same social computing program can be executed in experimental environments and in real-world interaction spaces. The demonstrator for the coordination of emergency response services may go live and coordinate human and nonhuman actors in genuine disaster recovery scenarios. Concerning its impact on the physical environment it possesses a virtual actuality in the test-bed environment and a real actuality when it is employed in real-time in order to control processes in the natural word.

In case of social computing systems the "actual sphere of realizing possible ends" can either be an experimental environment composed exclusively of software agents or a system running in real-time. In the latter case humans may be integrated for clarifying and/or deciding non-formalized conflicts in an ad-hoc manner. Automatic collaborative routines or new practises for ad-hoc collaboration are established. Novel purely virtual or hybrid contexts realizing collective and distributed agency materialize. Therefore it becomes vital to understand agency and interagency in virtual and hybrid constellations.

2 ATTRIBUTING AGENCY IN SOCIOTECHNICAL SYSTEMS

In order to exemplify the state of the art in attributing collective and distributed agency in sociotechnical systems two thought provoking schools are shortly summarized: the Actor Network Theory (ANT) and the sociotechnical approach of attributing distributed agency of Rammert and colleagues. Both intend to analyse constellations of collective inter-agency by attributing agency both to human and nonhuman actors but they differ in essential aspects.

The ANT approach introduces a flat concept of agency and a symmetrical ontology applicable both to human and nonhuman

[1] Carl von Linde Akademie, Technische Universität München, Munich, Germany, Email: sabine@thuermel.de

actors (e.g.[10]) whereas the distributed agency approach of Rammert et al. promotes a leveled and gradual concept of agency based on the "practical fiction of technologies in action" ([11], [12]).

2.1 The Actor Network Theory (ANT)

As a practitioner of science and technology studies and a true technograph Bruno Latour was the first to attribute agency and action both to humans and non-humans [13]. Together with colleagues as Michel Callon a symmetric vocabulary was developed which they deemed applicable both to humans and non-humans [14, p. 353]. This ontological symmetry led to a flat concept of agency where humans and nonhuman entities were declared equal. Observations gained in laboratories and field tests were described as so-called actor networks, heterogeneous collectives of humans and nonhuman entities, mediators and intermediaries. The Actor Network Theory regards innovation in technology and sciences as largely depending on whether the involved entities – may they be material or semiotic – succeed in forming (stable) associations. Such stabilizations can be inscribed in certain devices and thus demonstrate their power to influence the further scientific evolution [15]. All activity emanates from so called actants [10, pp. 54]. The activity of forming networks is named „translation"[10, p. 108]. Statements made about actants as agents of translation are snapshots in the process of realizing networks [16, p. 199]. The central empirical goal of the actor network theory consists in reconstructively opening up convergent and (temporarily) irreversible networks [16, p. 205]. Thus the ANT approach could more aptly be called a "sociology of translation", an "actant-rhyzome ontology" or a "sociology of innovation [10, p. 9]. However, it should be noted that Latour has quite a conventional, tool-oriented notion of technology [12]. This may be due to the fact that smart technology and agent systems are nowhere to be found in his studies.

2.2 Distributed agency and technology in action

It is important to Werner Rammert and Ingo Schulz-Schäffer under what conditions we can attribute agency and inter-agency to material entities and how to identify such entities as potential agents [11, p. 9]. Therefore a gradual concept of agency is developed in order to categorize potential agents regardless of their ontological status as machines, animals or human beings. Rammert is convinced that "it is not sufficient to only open up the black box of technology; it is also necessary and more informative to observe the different dimensions and levels of its performance" [12, p. 11]. The model is inspired by Anthony Giddens' stratification model of action [17]. It distinguishes between three levels of agency:

- causality ranging from short-time irritation to permanent re-structuring,
- contingency, i.e. the additional ability "to do otherwise", ranging from choosing pre-selected options to self-generated actions, and, in addition, on the highest level
- intentionality as a basis for rational and self-reflective behaviour [11, p. 26], [12, pp. 1].

The "reality of distributed and mediated agency" is demonstrated e.g. based on an intelligent air traffic system [12, p. 15]. Hybrid constellations of interacting humans, machines and programs are identified. Moreover a pragmatic classification scheme of technical objects depending on their activity levels is developed. This permits to classify the different levels of "technology in action". It starts with passive artifacts, continuing with reactive ones, i.e. systems with feedback loops. Next come active ones, then proactive ones, i.e. systems with self-activating programs. It ranges further up to co-operative systems, i.e. distributed and self-coordinating systems [18, p.7]. The degrees of freedom in modern technologies are constantly increasing. Therefore the relationship between humans and technical artifacts evolves "from a fixed instrumental relation to a flexible partnership"[12, p. 13]. Rammert identifies three types of inter-agency: "*interaction* between human actors, *intra-activity* between technical agents and *interactivity* between people and objects" [18, p. 8]. These capabilities do not unfold "ex nihilo" but "medias in res". "According to [this] concept of mediated and situated agency, agency arises in the context of interaction and can only be observed under conditions of interdependency" [12, p. 5].

These reflections show how „technology in action" may be classified and how constellations of collective inter-agency can be evaluated using a gradual and multi-level approach. Similar to Latour these authors are convinced that artifacts are not just effective means, but must be constantly activated via practise (enactment) [19, p. 15].

Since this approach focuses exclusively on „agency medias in res", i.e. on snapshots of distributed agency and action, the evolution of any individual capabilities, be they human or nonhuman, are not accounted for. Even relatively primitive cognitive activities as learning via trial and error, which many machines, animals and all humans are capable of, are not part of the methodical symmetry between human and technology. A clear distinction between human agency, i.e. intentional agents, and the technical agency, a mere pragmatic fiction, remains. In Rammert's view technical agency "emerges in real situations and not in written sentences. It is a practical fiction that has real consequences, not only theoretical ones" [12, p. 5]. In his somewhat vague view the agency of objects built by engineers "is a practical fiction that allows building, describing and understanding them adequately. It is not just an illusion, a metaphorical talk or a semiotic trick" [12, p .8].

3 LEVELS OF ABSTRACTIONS FOR SOCIAL COMPUTING SYSTEMS

In the following I want to base my approach on Rammert et al.'s reflections on the qualities of advanced technology in action. But in contrast to Rammert the agency of technology is not considered a "pragmatic fiction" but a level of abstraction (LoA**)**, as defined by Floridi. A pragmatic fiction is essentially a manner of speaking whereas a LoA corresponds to a (functional) abstraction. A LoA „is a specific set of typed variables, intuitively representable as an interface, which establishes the scope and type of data that will be available as a resource for the generation of information" [20, p. 36]. For a detailed definition see [21, pp. 44].

A LoA presents an interface where the observed behavior – either in virtual actuality or real actuality - may be interpreted. Under a LoA, different observations may result due to the fact that a social computing software can be executed in different

runtime environments, e.g. in a test-bed in contrast to a real-time environment. Different LoAs correspond to different abstractions of one and the same behaviour of social computing systems in a certain runtime environment. Different observations under one and the same LoA are possible if different versions of a social computing program are run. This is the case when software agents are replaced by humans.

Conceptual entities may also be interpreted at a chosen LoA. Note that different levels of abstraction may co-exist. Since levels of abstractions correspond to different perspectives, the system designer's LoA may be different from the sociologist's LoA or the legal engineer's LoA of one and the same social computing system. These LoAs are related but not necessarily identical.

The basis to technology in action is not a pragmatic fiction of action but a model of the desired behavior. From the designer's point of view metaphors often serve as a starting point to develop e.g. novel heuristics to solve NP-complete (optimization) problems or to build humanoid service robots instead of industrial robots. Such metaphors may be borrowed from biology, sociology or economics. Research areas as neural nets, swarm intelligence approaches and electronic auction procedures are products of such approaches. In the design phase ideas guiding the modeling phase are often quite vague at first. In due course their concretization results in a conceptual model [22, p. 107] which is then specified as a software system. From the user's or observer's point of view during runtime the more is known about the conceptual model the better its potential for (distributed) agency can be predicted and the better the hybrid constellations of (collective) action, emerging at runtime, may be analysed. Latour's snapshots are complemented by a perspective on the system model. The philosophical value added of this approach does not only lie in a reconstructive approach as intended by Latour and Rammert but also in the conceptual engineering of the activity space. Under a LoA for agency and action, activities may be observed as they unfold. Moreover the system may be analysed and educated guesses about its future behaviour can be made. Both the specifics of distinct systems and their commonalities may be compiled.

4 MULTIDIMENSIONAL GRADUAL AGENCY

The following proposal for a conceptual framework for agency and action is intended to provide a multidimensional gradual classification scheme for the observation and interpretation of scenarios where humans and nonhumans interact. It permits to define appropriate lenses, i.e. levels of abstraction, under which to observe, interpret, analyse and judge their activities.

As Rammert states, "agency really is built into technology" but – in my opinion - not "as it is built into people" [12, p.6] but by intelligent design performed by engineers and computer scientists. In order to demonstrate the potential for agency not only the activity levels of any entities but also their potential for adaptivity, interaction, personification of others, individual action and conjoint action has to be taken into account. Being at least (re)active is the minimal requirement for being an agent. Higher activity levels permit to influence the environment. Being able to adapt is a gradual faculty. It starts with primitive adaption to environment changes and ranges up to the adaption of long-term strategies and the corresponding goals based on past experiences and (self-reflective) reasoning of human beings. Based on activity levels and on being able to adapt in a "smart" way acting may be discerned from just behaving.

The potential for interaction is a precondition to any collaborative performance. The potential of the personification of others enables agents to integrate predicted effects of own and other actions. „Personification of non-humans is best understood as a strategy of dealing with the uncertainty about the identity of the other …Personifying other non-humans is a social reality today and a political necessity for the future" [23, p. 497]. It starts with the attribution of simple dispositions up to perceiving the other as a human-like actor. This capability may affect any tactically or strategically motivated individual action. Moreover it is prerequisite to any form of defining conjoint goals and conjoint (intentional) commitment. The capabilities for individual action and conjoint action may be defined based on activity levels, the potential for adaptivity, interaction and personification of others possessed by the involved actor(s).

Any object entity type may be classified according to its characteristics in these dimensions. For any entity types the maximum potential (in these dimensions) is defined by a distinct value tuple. It may be depicted by a point in the multidimensional space spanned by the dimensions introduced above.

Any token, i.e. instantiation of an entity type, may be characterized by a distinct value tuple at a moment in time, i.e. by its actual time-stamped value. This value reflects the virtual actual activity if the program is run in a test-bed. It portrays its real actuality if the program is run in real-time in a real world environment. In agent-based systems the changes over time correspond to state changes of each agent.

Note that in the following the granularity on the different axes is only exemplary and can be adjusted according to the systems to be analysed and/or compared.

The activity level permits to characterize individual behaviour depending on the degree of self-inducible activity potential. It starts with passive entities as Latour's well-known road bumpers. Reactivity, realized as simple feedback loops or other situated reactions, is the next level. Active entities permit individual selection between alternatives resulting in changes in the behavior. Pro-active ones allow self-reflective individual selection. The next level corresponds to the capability of setting one's own goals and pursuing them. These capabilities depend on an entity-internal system for information processing linking input to output. In the case of humans it equals a cognitive system connecting perception and action. For material artifacts or software agents an artificial "cognitive" system couples (sensor) input with (actuator) output.

Based on such a system for (agent-internal) information processing the level of adaptivity may be defined. It characterizes the plasticity of the phenotype, i.e. the ability to change one's observable characteristics including any traits, which may be made visible by a technical procedure, in correspondence to changes in the environment. Models of adaptivity and their corresponding realizations range from totally rigid to simple conditioning up to impressive cognitive agency, i.e. the capability to learn from past experiences and to plan and act accordingly. A wide range of models co-exist allowing to study and experiment with artificial "cognition in action". This dimension is important to all who define agency as situation-appropriate behavior and who deem the plasticity of the phenotype as an essential assumption of the conception of man.

The potential for interaction, i.e. the coordination by means of communications is the basis to most if not all social computing systems and approaches to distributed problem solving. It may range from uncommunicative to hard-wired cooperation mechanisms up to ad-hoc cooperation.

The personification of others lays the foundation for interactive planning, sharing strategies and for adapting actions. This capability is non-existent in most material and software agents. Some agents have more or less crude models of others, e.g. realized as so-called minimal models of the mind. A next qualitative level may be found in great apes [24] which also have the potential for joint intentionality. This provides the basis for topic-focused group decision making based on egoistical behavior. Understanding the other as an intentional agent allows even infants to participate in so-called shared actions [25]. Understanding others as mental actors lays the basis for interacting intentionally and acting collectively [25]. Currently there is quite a gap between nonhuman actors and human ones concerning their ability to interact intentionally. This strongly limits the scope of social computing systems when it is used to predict human behavior or if it is intended to engineer and simulate future environments.

Both the potential for individual action and for conjoint action may be defined based on the above mentioned capabilities for activity, adaptivity, interaction and personification of others. One option is the following: In order to stress the communalities between human and nonhuman agents, an agent counts as capable of acting (instead of just behaving), if the following conditions concerning its ontogenesis hold: "the individual actor [evolves] as a complex, adaptive system (CAS), which is capable of rule based information processing and based on that able to solve problems by way of adaptive behavior in a dynamic process of constitution and emergence" [26, p. 320]. Based on the actor's capability for joint intentionality resp. understanding the other as an intentional agent or even as a mental actor, the actor may be able of joint action, shared or collective action in the sense outlined above. New capabilities may emerge over time on the individual level (e.g. emergent semantics, emergent consciousness). Self-organisation and coalition forming on the group level can occur. New cultural practices and novel institutional policies may emerge.

Constellations of inter-agency and distributed agency in social computing systems or hybrid constellations, where humans, machines and programs interact, may be described, examined and analysed using above introduced classification scheme for agency and action. These constellations start with purely virtual systems like swam intelligence systems and fixed instrumental relationships between humans and assistive software agents where certain tasks are delegated to artificial agents. They continue with flexible partnerships between humans and software agents. They range up to loosely coupled complex adaptive systems. The latter may model so diverse problem spaces as predator-prey relationships of natural ecologies, legal engineering scenarios or disaster recovery systems. Their common ground and their differences may be discovered when the above outlined multi-dimensional, gradual conceptual framework for agency and action is applied. A subset of these social computing systems, namely those which may form part of the infrastructure of our world, provide a new form of "embedded governance". Their potential and limits may also be analysed using the multi-dimensional agency concept.

5 CONCLUSIONS & FUTURE WORK

The proposed conceptual framework for agency and action offers a multidimensional gradual classification scheme for the observation and interpretation of scenarios where humans and nonhumans interact. It may be applied to the analysis of the potential of social computing systems and their virtual and real actualizations. The above introduces approach may also be used to describe situations, where options to act are delegated to technical agents. The corresponding variants of e-trust and potential legal relationships may be characterized.

REFERENCES

[1] N. Jennings. *ALADDIN End of Project Report*, http://www.aladdinproject.org/wp-content/uploads/2011/02/finalreport.pdf (2011), accessed January 24th, 2012

[2] *K. Mainzer . Thinking in Complexity. The Complex Dynamics of Matter, Mind, and Mankind*, Springer Verlag: Berlin, Germany, 5th edition, (2007).

[3] M. Wooldridge. *An Introduction to Multi-Agent Systems*, John Wiley & Sons Ltd: England, UK, (2002).

[4] T. Bosse, A. Sharpanskykh and J. Treur. Integrating Agent Models and Dynamical Systems. In: Declarative Agent Languages and Technologies V, LNCS 4897, Springer: Heidelberg, 50-68, (2008).

[5] Economist Print Edition. *Agents of Change*, 22-Jul-2010, online http://www.economist.com/node/16636121/print. 2010, (2010), accessed January 24th, 2012

[6] Leibnitzcenter for Law. Multi-agent PhD position availabe. www.leibnitz.org/wp-content/uploads/2011/02/wervering.pdf, accessed January 24th, 2012

[7] D. Berthier. *Médiations sur le réel et le virtuel*, Editions L'Harmattan : Paris, France, (2004).

[8] D. Berthier. *Qu'est-ce que le virtuel.* http://www-lor.int-evry.fr/~berthier/ JR-Qu-est-ce-que-le-virtuel.pdf (2007), accessed January 24th, 2012

[9] Chr. Hubig. *Die Kunst des Möglichen I – Technikphilosophie als Reflexion der Medialität*. Transcript Verlag: Bielefeld, Germany, (2006).

[10] B. Latour. *Reassembling the Social – An Introduction to Actor-Network-Theory*, Oxford University Press: Oxford, U.K, (2005).

[11] W. Rammert and I. Schulz-Schäffer. Technik und Handeln: Wenn soziales Handeln sich auf menschliches Verhalten und technische Abläufe verteilt. In: *Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik, W.* Rammert and I. Schulz-Schäffer (eds), Campus: Frankfurt, Germany, 11-64, (2002).

[12] W. Rammert. *Distributed Agency and Advanced Technology Or: How to Analyse Constellations of Collective Inter-Agency***,** Berlin: The Technical University Technology Studies Working Papers TUTS-WP-3-2011, (2011).

[13] B. Latour . Mixing Humans and Nonhumans Together. The Sociology of a Door-Closer. In: Social Problems, Vol 35, No 4, 298-310, (1988).

[14] M. Callon and B. Latour. Don't throw the baby out with the bath school" A reply to Collins and Yearley. In: *Science as Practise and Culture*, A. Pickering (ed), University of Chicago Press: Chicago, U.S., 343-368, (1992).

[15] B. Latour. Drawing Things Together. In: *Representation in Scientific Practice*. M. Lynch and St. Woolgar (eds), MIT Press: Cambridge, Mass, U.S., 19-68, (1990).

[16] I. Schulz-Schaeffer, Ingo „Akteur-Netzwerk-Theorie. Zur Koevolution von Gesellschaft, Natur und Technik", in: Weyer, Johannes (Hrsg.): *Soziale Netzwerke. Konzepte und Methoden der sozialwissenschaftlichen Netzwerkforschung. R. Oldenburg Verlag: München, Germany*, 187-209, (2000).

[17] A. Giddens. *The Constitution of Society, Outline of the Theory of Structuration*. Polity Press: Cambridge, UK, (1984).
[18] W. Rammert. Where the Action is: Distributed Agency between Humans, Machines and Programs. In: *Paradoxes of Interactivity*, U. Seifert, J. H. Kim and A. Moore (eds). Transcript and Transaction Publishers: Bielefeld and New Brunswick, Germany and U.S., 62-91, (2008).
[19] W. Rammert. *Die Techniken der Gesellschaft: in Aktion, in Interaktivität und in hybriden Konstellationen*, Berlin: The Technical University Technology Studies Working Papers TUTS-WP-4-2007, (2007).
[20] L. Floridi. *The Method of Levels of Abstraction*, Minds and Machines, 2008, vol. 18, No 33, 303-329, (2008).
[21] L. Floridi. *The Philosophy of Information*, Oxford University Press: Oxford, UK, (2011).
[22] A. Ruß, D. Müller and W. Hesse. Metaphern für die Informatik und aus der Informatik. In: Menschenbilder und Metaphern im Informationszeitalter. M. Bölker, M. Gutmann and W. Hesse (eds), LIT Verlag: Berlin, Germany, 103-128, (2010).
[23] G. Teubner. Rights of Non-humans? Electronic Agents and Animals as New Actors. In: *Politics and Law* (Journal of Law & Society 33), 497-521, (2006).
[24] J. Call and M. Tomasello. *Does the chimpanzee have a theory of mind? 30 years later*, Trends in Cognitive Science, 12, 187-192, (2008).
[25] M. Tomasello. *Origins of Human Communication*. MIT Press: Cambridge, U.S., (2008).
[26] P. Kappelhoff . Emergenz und Konstitution in Mehrebenenselektionsmodellen. In: J. Greve and A. Schnabel (eds.) *Emergenz – Zur Analyse und Erklärung komplexer Strukturen*, suhrkamp taschenbuch wissenschaft 1917, Suhrkamp Verlag: Berlin, 319-345, (2011).

Collective Individuation: A New Theoretical Foundation for post-Facebook Social Networks

Yuk Hui[1], Harry Halpin[2]

Abstract. Despite their increasing ubiquity, there is no fundamental philosophical theory of social networking, and we believe this has limited the technical development social networking to very limited use-cases. We propose to develop a theoretical discourse on the new generation of social networks and to develop software prototypes for an alternative. Our project centres on the question: what is collective individuation and what is its relation to collective intelligence? Current social networking websites and network-science are based on individuals as the basic analytic unit, with social relationships as simple "ties" between individuals. In contrast, this project wants to approach even individual humans as fundamentally shaped by their collective social relationships, building from Simondon's insight that individuation is always simultaneously psychological and collective. Our proposal should enable new kinds of social imagination and social structure through redesigning the concept of the 'social' in the time of Facebook.

1 FACEBOOK AND THE PROBLEM OF INDIVIDUATION

a) The Origin of Social Networks: Moreno and Saint Simon

One of the emerging research areas of web science and network analysis is the attempt to analyze social networks in terms of network theory as it directly descends from sociological approach by questionnaires, interviews which attempt to understand the social relations and explain certain social phenomenon. The marriage of this sociological approach and mathematical representations during the early-mid 20th century gave us a significant image to think about the 'social', in which individuals are often considered as nodes and their social relationships are mapped to edges. This pioneered the application of graph theory in social network analysis. Today with the assistance of computers which facilitate data collection and image processing and especially the rise of social networking website, such a conceptualization seems to be a foundation of a new discipline mediating the computer science and sociology and cultural studies. In its entirety, the image of network consisting of nodes becomes the representation and also a method to approach social phenomenon. To us, the problem is that this approach takes for granted many historical developments and philosophical assumptions. Our questions start from: *where did this entire conception come from? What legitimates its being? What is the consequence of such a conceptualization?* These questions constitute the first part of this article; in the second part, we will propose another way to think of social networks and discuss the alternatives.

J. L. Moreno(1889-1974), a psychologist and founder of sociometry was one of the first sociologists to demonstrate the value of graph-theoretic approaches to social relationships. The most-often quoted example is Moreno's work at the New York State Training School for Girls in Hudson where the run-away rate of the girls were 14 times more than the norm! Moreno identified it as a consequence of the particular network of social relationships amongst the girls in the school, and he followed by creating a simple sociological survey to help him to "map the network". The survey consists of simple questions such as 'who do you want to sit next to?' Moreno found from the map that the actual allocation plan of the girls in different dormitories created conflicts; he then used the self-same model to propose another allocation plan that successfully reduced the number of run-away. The belief in the representation of social relations by 'charting' prompts Moreno to write that 'as the pattern of the social universe is not visible to us, it is made visible through charting. Therefore the sociometric chart is the more useful the more accurately and realistically it portrays the relations discovered.' [1] But one should be careful that by doing this, the charting is no longer a mere representation of social relationships, but also that these maps of social relationships could be used to realize what Moreno called social planning, meaning to reorganize "organic" social relationships with the help of planned and technologically-embodied social networks. At this point that we can identify a question which is not yet been tackled significantly by researches, which Moreno already proposed in 1941: the superimposition of technical social networks upon pre-existing social networks 'produces a situation that takes society unaware and removes it more and more from human control' [2] This lost of control is the central problem of the technical social networks currently, and in order to address this phenomenon, we propose to question some of the presuppositions that have been hidden in the historical development of social network analysis.

Despite their explicit mapping of social relationships, social networking analysis is actually an extreme expression of social atomism. This proposition has to be understood sociologically and philosophically: The presupposition of the social networks is that individuals constitute the network, and hence individuals – which in traditional sociology (if we count Actor Network Theory as an alternative), tend to be humans - are the basic unchanging units of the social networks. If there is any collectivity, it is considered primarily being created by the sum of the individuals and their social relationships as quantifiable

[1]Institut de Recherche et d'Innovation du Centre Pompidou, Paris, www.digitalmilieu.net/yuk
[2]World Wide Web Consortium, MIT, http://www.ibiblio.org/hhalpin/

representation in the map of the networks. This view is at odds with what has been widely understood in anthropology: namely that a society, community, or some other collectivity are beyond the mere sum of individuals and their relationships. It can be noted that historically the development of collectives has originally existed in the form of families, clans, tribes, and so on and so forth even pre-dates the notion of the autonomous individual[3].

The reemergence of sociometry should attribute to the proliferation of technical networks, and here we must recognize that today is not longer human relations are mapped in sociometry but virtually anything which can be digitalized, or more precisely anything can be represented as data and relations can be established according to two different terms. The arrival of network society supported by technological infrastructure further reinforces the concept of sociometry. Lets recall that in 1933 when Moreno published in New York Time an article 'Emotion Mapped' where he suggested to draw a sociometric map of New York City, in fact he could only work on community of size 435, nowadays with tools such as Facebook, Moreno's dream is not impossible[3]. At the same time, the combination of the social and the network also reactivates the spirit of industrialization which one can trace back to the 19th century French philosopher, socialist Saint Simon. The French sociologist Pierre Musso shows that Saint-Simon was the first philosopher who fully conceptualized the idea of networks via his understanding of physiology, which he then used to analyze vastly different domains, albeit more imaginatively rather than concretely as done later by Moreno.[4] Saint Simon indeed envisioned networks as including communication, transportation, and the like, holding the idea of a network as both his primary concept and tool for social transformation. Saint Simon believes that through industrialization, it is possible to create a socialist state by reallocating wealth and resources from the rich to the poor, from the talented to the less talented, like an organism attains its inner equilibrium by unblocking all the circulations.

Today we know from history that Saint Simon's sociology was blind to the question of classes which was later analyzed by Karl Marx in *Das Kapital*. Marx's vision of the society is often distorted as social planning, which is more or less the codification of collections in the Soviet fashion. Moreno criticized this distorted figure of Marx and proposed that the 'next social revolution will be of the "sociometric" type. The revolutions of the socialistic-marxistic type are outmoded ; they failed to meet with the sociodynamics of the world situation'. Moreno's announcement maybe demonstrated today by Facebook as some of the pop writers on technology would say, but in fact what Moreno means by that has to be further discussed, especially the concept of spontaneity. But neither Saint Simon's distinctly old-fashioned industrial vision is considered, since it is obviously that socialism doesn't come naturally through industrialization, but what is new is the imagination of a new democratic society, which is frictionless through the mediation of networks. By frictionless here we mean the conceptualization a rather flattened social structure with kind of slogans such as 'Here Comes Everybody'; one can use Facebook and etc to autonomously organize events, movements, and even revolutions. It is the same for Moreno, the sociometric revolution never gets rid of its own shadow.

b) Alienation and Disindividuation

The graphical portrayal of social networks as nodes and lines reinforces the perception of Moreno and Saint Simon that social relations always exit in the form from one atomic unit to another. This *image,* with its obvious bias towards vision[4], has become the central paradigm in understanding society and the technological systems. Yet any image is also a mediation between the subject and object that pre-configures – or pre-programs – a certain intuition onto the world[5]. One can imagine that the image itself of a social network as merely lines and dots constrains innovation as it cannot understand how to graphically represent any collectivity beyond the individual as primacy, but always take it only consequence or byproduct of the map of interconnected atoms. This is something Moreno forgot or he couldn't see at his time: the materialization of social relations, not in the figure of charts on the paper, but controllable data stored on the computer which mediate the actions of users. What Moreno called a sociometric revolution is a postulation that through certain sociometric planning, the spontaneity of human interactions can be enhanced. Moreno gained this insight from his long time works on psychodrama, based on which he criticized psychoanalyst especially Freud couldn't 'act out'. What Moreno means by 'acting out' in this context is that the psychoanalysts feared to participate in the theater of the patient, but only act as a mere observer. We want to add more meanings to this word 'acting out' in the passages followed. But here we want to point out that firstly seeing each individual as a social atom already implies an extreme form of individualism that intrinsically dismisses the position of the collective; secondly today when sociometrical vision is materialized in social networking website, what is at stake is exactly Moreno's own faith in spontaneity and the question of individuation.

Social networking sites like Facebook stay within this paradigm by providing only digital representations of social relations that pre-exist in a richer social space, and allows new associations based on different discovery algorithms to emerge. Facebook's very existence relies largely on the presupposition of

[3] Such a view of individualism is also naturalized in economic studies since Adam Smith, who saw division of labour as a natural development and the exchange between individuals as the origin of economic life. In the works of anthropologists such as Marcel Mauss, David Graber, we can find another understanding of economy which is since the beginning collective.

[4] It has been widely criticized in the 20th century that western philosophy has a bias towards vision, we see this in the work of Heidegger and etc. It is interesting to note that Guy Debord even criticized it as a weakness 'The spectacle inherits the weakness of the Western philosophical project, which attempted to understand activity by means of the categories of vision, and it is based on the relentless development of the particular technical rationality that grew out of that form of thought.', see Guy Debord, The Society of Spectacles, §19, Chapter 1, http://www.bopsecrets.org/SI/debord/1.htm

[5] One can also speak of the *Weltbild* as deployed by Heidegger, where Heidegger showed that an image is not simply a representation of the world, but also that the world can be controlled and manipulated as an image.

individualism, as the primary unit in Facebook is always the individual's Facebook profile. One can always recall the original idea of Facebook, as it was shown in the film, the young Mark Zuckerberg created Facebook as a tool to express his sexual desire, that is to say a libidinal economy intrinsically individualistic. This exploitation of libidinal economy is not new today, in the past decades, we already witnessed the exploitation of libidinal energy in consumerism[6]. In the turn of the 20th century, the father of public relations, Edward Bernays adopted psychoanalysis in his marketing techniques and integrated the economy of commodities with the libidinal economy. It may be interesting to note that in fact Bernays is the nephew of Sigmund Freud.

Bernays employed the psychoanalysts to participate in designing marketing strategies. One of the well known examples is to promote the tobacco business to the American females, since at that time the female smoking population in the United State is quite low. Bernays hired the female movie stars to smoke in the public, this create a circuit of libidinal economy which has to be completed through the action of smoking, which is also to say buying the cigarette. Today it is no longer simply cigarettes, but whatever commodities. Here is the picture of the consumerism of the 20th century: the workers sell their labour-time to the factories and offices, afterwards they are seduced to spend their salaries on the unnecessary and magical commodities – the control of both physiological and psychological circuit. On Facebook, it seems as if the users have their own will to execute actions, but in such as technological system, the vision, actions have to adopt the configurations and functions of the system. In general, on other sites such as Google+ group profiles or anonymous profiles are actively discouraged. One cannot deny that these social networks are able to bring people together and form groups whose activity ranges from shopping to protests. Yet we have to be careful here, as these groups are positive externalities in economic terms. These social networking website support only a few collective actions, but are instead optimized for individuals to map their own network of friends so they can leave individuals commenting on each other's posts and clicking on very basic individual operations such as 'Like' and 'Want', which are now increasingly littered throughout the entire Web.

When the users are considered as social atoms which can then be superimposed onto a technological network, the spontaneity and innovation within the collective is given to control of the networks, which is mainly driven by intensive marketing and consumerism aimed at individuals[7]. Social networks have obviously become both an apparatus to express and control the desire of the users. The subject is an atom, and within the social networks, subjectivation becomes an engineering process subjected to careful monitoring and control, which has been thought of by theorists like François Perroux[8] as a source of a new kind of alienation. This is not entirely dissimilar to the alienation which Marx described in *Das Kapital* which was produced by having human workers adapt to the rhythm of the machines, so the worker loses control of his vital energy and ultimately his time to reflect and to act. When Marx describes the vital forces of the collective, he uses the German word *Naturwüchsigkeit*, which can literally translated into English as the nature-growth-ness, which is similar to what Moreno calls spontaneity[9]. The similarity lies in the imagination of the autonomous subjects naturally interact with each other and create a collective that at the same time displaces the individuals. And Moreno's 'acting out' as a psychologist is also the catalyst for the 'acting out' of the collective. The second sense of the acting out is the formation of group conditioned by a projects, it designates an investment of attention; libidinal energy and time. If an existential critique can be introduced here, we can say time and equally the attention of each social atom is chopped into smaller pieces and disperse on the networks by the status updates, interactions, advertisements, and the like. This form of collective that is exactly what Martin Heidegger would call 'das Man', the 'they' who exhausts one's time without giving meaning to one's own existence. In fact, Bernard Stielger would hold that these constructed social atoms are not individuals are not really 'individuals', but the disindividuals, as they seem to have lost their ability to act out and to relate except within the apparatus of an atomistic social network[10]. [5]

c) Social Engineering and Technical Engineering

Moreno's sociometry as response to both Marx' economic materialism and Freud's psychological materialism encounters its own impasse today; Moreno and Saint-Simon didn't take digital networks and telecommunication into account in their theories – yet nonetheless technological materialism is currently tied to this new digital economic, psychological, and technological network.[6] Society is mediated by data. Sites like Facebook uses graphs of personal connections to predict and hence 'recommend' products, and so produce desires in the individual that show that the autonomous individual is in fact shaped not only by their relationships in the network, but by the existence of the network itself. While the Internet is a distributed and decentralized network, industrialization reverses this principle as simply to maintain a social graph for analysis the

[6] Bernard Stiegler, For a New Critique of Political Economy, Polity, London, 2010

[7] After the Like button, Facebook has announced in September 2011 of introducing the Want button, that is designed for marketing, http://www.auctionbytes.com/cab/abn/y11/m09/i23/s01

[8] The French economist François Perroux took up the question of industry and social transformation from Saint-Simon and developed a vision of collective creation, in which humans and machines act on each other and through the standardization of objects, human beings can renew their life style, and produce a system of 'auto collective creation'. Notably Perroux was also influenced by Schumpeter, especially the concept of creative destruction.

[9] Hence one should recognize the problematic of Moreno's critique of Marx, and one may be able to develop a new relation between Moreno and Marx

[10] B. Stiegler, états de choc : Bêtise et savoir au XXIe Siècle, Mille et une Nuit, 2012, p.102-105, where he proposes three types of disindividuation, firstly the regression to the pure social, what is pure social is the animal form of life; secondly the deskilling process by technologies, for example when the craftsmen had to enter factories and gave up their own skills and way of life; thirdly the process of 'bracketing' the previous individuation which produces a 'quantum jump' and exceed the threshold of the psychical transformation, according to Stiegler, these three types of disindividuations cannot be separated.

size of Facebook requires immense centralization. At the same time it creates a *technical reality*, with a deception of being an unmodifiable default. Yet, we have to ask: is Facebook a social collectivity, or the false image of one? Going beyond the social graph, we need to grasp other possibilities of 'social networks'.

The social engineering of facebook is supported by its multiple features ranging from sharing and 'I like' functions to privacy settings. Here we sees the unification of social engineering and technical engineering, which also poses the great challenge to the humanities. It will be necessary to look at how these realities are created and accepted, for example if one tries to leave, one losses everything, including the social relations, profile data, the possibility of communicating with friends. Even when one uses social networking sites, individuals and expressions are conditioned by the capacities permitted according to the features of the website and there is little to no privacy. One cannot choose to be anonymous, on the other hand the verification of identities become more and more an important to industry.

There can be political considerations, for example, in China the social networks request the users to prove their identities by showing their identity cards, and this may be in response to the fact that the question of anonymity is seemingly increasingly important for democracy and transparency as has been shown by Wikileaks. There is even a demand for anonymity, as the Japanese Ni Chanel(2ch) which entirely operates on the basis of anonymity has became one of the most popular social network website in Japan. These features would obviously be vital to those in the Middle East, London, Spain, and #OccupyWallSt. If subjectivation within social networks is an engineering process, what is necessary is to produce a new type of thinking and new form of social networks. Some of this thinking can be seen in various slogans: data portability, privacy and personal possession of data. These slogans are natural responses to the monstrous ability of social networks to create "walled gardens" out of personal data. Though these slogan are important to fight against the dictatorship of Facebook, they still lack an overall reevaluation of facebook and a vision of an alternative social network which is not merely an immediate response.

2 PROJECT, PROJECTION AND COLLECTIVE INDIVIDUATION

a) Simondon and Collective Individuation

Hence we propose to rethink from the perspective of the collective, as a remedy to the individualistic approach of the current social networks. This doesn't mean they we want simply collectivity, but rather we want to put collective at the same level as individual, like water and fish which cannot be vivant without each other. Sociometry demands a mapping which is becoming more and more precise, and reflects the probabilities of connections, interactions, marketing, that is a technological individuation easily slips back to *disindividuation*. Can we think of an new kind of individuation that cannot be reduced to statistics, and whose power only work in ambiguity, instead of precisions? We propose that the French philosopher Gilbert Simondon proposed in his book *L'Individuation psychique et Collective* a model of individuation which can be therapeutic to the conceptualization of the social presupposed by the current technological developments- or in other words socio-techno engineering.[7]

Simondon suggests that individuation is always both psychical and collective. What Simondon means by psychical individuation can be considered to be the psychology of individuals, for example under the situation of anxiety, grief, angry, etc. But pure psychic and pure social are not enough. For Simondon, individuals and groups are not opposite to each other, meaning while in the group, one loses his or her singularity, as what was considered as the Soviet type of collectivism. Instead, the individual and the group constitute a constant process of individuation. Psychical individuation to Simondon is more an individualization, which is also the condition of individuation, while collective individuation is one that brings the individual to constant transformation. Hence one can understand that nature is in fact not in opposition to human being, but rather the primary phase of being, human being and the technical milieu created by them constitute the second phase of being, which if we can say so, it is the technical individuation proposed by Bernard Stiegler.

Simondon hence rejected the American microsociology and psychology, which indirectly includes Moreno's sociometry (via the works of Kurt Lewin), as being substantialism. The substantial approach towards individuals and groups easily ignores the dynamic of the social, and see individual and collective as interiority and exteriority that has to be separated . This approach falls prey to the extreme of psychologism and sociologism – a molecular and molar substantialism- which consider individuals precede groups or groups precede individuals. The former sees the psychology of the individuals as the determining factor of the collective, and consider the formation of the collective only by considering: why the individual wants to participate- a typical question for those who do marketing or planning a start-up; the later sees social norms and collectives as predefined structures, that is to say in order to form a collective one needs immediately set up the social categories and 'mould' the individuals according to these pre-configurations.

Simondon considers individuation as a process of crystallization. Considering a supersaturated solution is undergoing crystallization, by absorbing energy each individual ion is transforming itself according to the relations with others, that is its milieu. It is the same in the group genesis that each individual is at the same time agent and milieu.In contrast, crystallization is a process that though finally gives a form, e.g, the identity of a specific crystal, it is also at the same time a process depends less on the form(on can always figure out forms) but rather on the redistribution of energy and matter. Simondon hence proposes to think of individuation as a necessary dynamics between individuals and groups. He distinguishes 'in group' and 'out group', and suggests to think of 'in group' as an intermediate between individual beings and 'out group'. One may sense a bit of similarity between Moreno and Simondon in this respect, that is the spontaneity of in-group and out-group; and it is also by this reason that we believe Moreno's sociometric technique though can be used today to analyse social networks like Facebook, Twitter, but it also post tremendous danger of social engineering that fall back to psychologism and

sociologism if we ignore his discussion on spontaneity, while we won't be able to fully discuss it in this short article.

b) Projects as the Basic Unit of Group

One may want to ask: isn't what we have seen on Facebook already a psychic and collective individuation? It is true that the philosophical approaches of Simondon can become tools to analyze social relations, but one must go beyond the limit that thoughts are merely tools of analysis, and recognize that they are also tools for transformation. As we have seen, Facebook individuates primarily atomistic individuals, and we propose to start from the collective instead in order to redesign the relation between the individual and the collective. Instead of how social atoms form collective, we must find out how a collective social network changes and shapes the individuals, and take this phenomenon as primacy. This social network will be one that enables collective individuation but also as a remedy to the industrial intoxication and exploitation of libidinal energy.

Hence we want to reflect on the question of group, and we want to propose that what distinguishes a collective from an individual is the question of a common project pertaining to groups. Take for example Ushahidi, a website that provided mapping service. After the earthquake in Haiti in 2010, in order to help Haiti to recover from the catastrophe. By using a web-based platform, Ushahdidi enabled both local and overseas volunteers to collect SMS messages with a special hash code to map the crisis in order help save people who might otherwise be lost. After the earthquake and tsunami in Japan in 2011, engineers from Japan developed a map of the damages caused by the tsunami and the emergencies need to be taken care of by analyzing tweets and other social medias. The dynamics of these projects go far beyond simply posting status updates, but allow people to dynamically work together on common goals. It is the moment of the formation of projects that allows the individuals to individuate themselves through the collective, and so give meaning to the individuals. On Facebook, one can establish a group, a page, an event, it seems to allow a common project to appear, but it doesn't provide the tools for collective individuation based on collaboration; in other words, on Facebook a group is no different from an individual.

Passing from a philosophical model to its realization in a technical system, we propose that the social networking site should exist as a set of tools to enable the collective creation and administration of a project. The collective intelligence is activated insofar as the group successfully uses its human and technical abilities to accomplish its goals. A user must always belong to a project, without which he or she will *not* be able to fully utilize the features – and projects are defined by groups. This is a first attempt to tackle the individualism exist in the current paradigm of social networks. Each project is defined by a goal and requirements of fulfillments as collectively initiated and updated by members of the group. Tasks will be assigned to users either in the form of individuals or subgroups, the progress of the tasks will be monitored and indicated. However, the collective should be dynamic rather than static, groups can be merged together to form larger projects and a project can also be split into smaller collectives. Groups can discover each other and communicate to seek possibility of collaborations and information sharing.

c) Case Studies and a Possible Framework

In our project 'Social Web', we look at some of the current models, including Wikipedia, some open source platforms, and alternative social networking projects like Lorea[11], Federated General Assembly[12], Crabgrass[13], and Diaspora - as well as unusual social networking websites such as Ni Channel, NicoNico Douga in Japan. Some of these groups already demonstrate the value of groups and projects, for example the encyclopedia project of Wikipedia, also Lorea and Crabgrass to create an alternative social networks that favor groups and common working spaces. We also recognize that though each of them has some of the collaborative features necessary for a new kind of social network, they don't really take the idea of individuation at the core of their designs. They can easily become examples of successful crowd sourcing that lows production cost and raising profits, instead of allowing us to rethink alternatives with different values and assumptions. Besides of returning to the primacy of groups, and emphasize on group management, we also suggest some other technical features for such a vision of collective social network:

1) The network primarily exists as directed social communication aiming at project, it also needs various other collaboration tools such as forums, wikis, etc. However, unlike traditional social networks, the purpose of the social networking site will be to help users store and refine data, the data can be stored in an open format such as RDF. Users and groups have the permission to manage data of the projects, and retrieve data using tagging and semantic search. Mapping should be employed as one possible, and easily interpretable, way to understand collective data collection.

2) Anonymity can be allowed under certain conditions (for example the group is wholly anonymous, or the group decides to open to anonymity) by collective projects. For example, in Ni Channel, one of the reasons that the inventor wanted it to be anonymous is that there won't be segregation between experienced users and amateurs, that might harm the formation of the collectives. [8] Besides of the possibility to yield interesting social phenomenon, anonymity can also act as a counter-force of the strict control of identities and censorship.

3) Personal data should be accessible only to the collective, and not even to those that run the server. Concerning the security of the networks, data either on the servers will be encrypted by implementing public key infrastructure, with the group being defined by shared public keys. Hence the ISP and system administrators won't be able to access the data on the server. Secondly the data will be stored distributed across multiple servers in order to minimize the consequences of attacks.

3 CONCLUSIONS & FUTURE WORK

[11] https://n-1.cc/pg/groups/7826/lorea/
[12] http://projects.occupy.net/
[13] https://we.riseup.net/crabgrass/about

The above outline is an introduction to a philosophical framework of a funded project titled ‘social web’. Facebook to us, represents an industrialization of social relationships to the extreme that it transforms the ‘social’ to a totally ‘atomic’ individualism. Saint Simon’s imagination of socialism based on the believe of the common good and well being of individuals through building networks is deemed to be a failure, but the relation between network and society take a more aggressive form at the time of ubiquitous metadata. Moreno’s sociometry technique probably finds its best companion today on Facebook and other social networking apparatus, but celebrating the reemergence of sociometric technique is only blind to the danger posed by the presuppositions of such theory and the technological developments that never examine its origins. We propose that social computing today must go beyond the traditional digital humanities, which proposes to analyze the social transformation by taking technologies into account, rather it will be more fruitful to follow what Stiegler calls pharmacology, which is to say technology is both good and bad, both a remedy and a poison at the same time, but it is necessary to develop a therapeutic approach against the toxicity generated by it, which in our case is Facebook(s).

Collective individuation proposes that another social network is possible, and it is necessary to consider an economy which is far more than marketing, click rate, number of users, etc. For us, a project is also a projection, that is the anticipation of a common future of the groups. By tiring groups to projects, we want to propose that individuation is also always a temporal and existential process, rather than merely social and psychological. By projecting a common will to a project, it produces a co-individuation of groups and individuals. The project is under development, but we hope the above outlines show the problem of the social networks and the limits of digital humanities (especially those who embraces sociometry) in understanding social computing, and it is clear that a new method towards software development is possible, and urgent.

REFERENCES

[1]J.L. Moreno, Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama, Beacon House Inc .Beacon, N. Y. 1978

[2] J. L. Moreno, Foundations of Sociometry: An Introduction, in sociometry, American Sociological Association , Vol. 4, No. 1 (Feb., 1941), pp. 15-35

[3] S. Wasserman and K. Faust, Social Network Analysis : Methods and Applications, New York [etc.] : Cambridge University Press, 1994

[4] P. Musso, Aux origines du concept moderne : corps et réseau dans la philosophie de Saint Simon. In: Quaderni. N. 3, Hiver 87/88. pp. 11-29. doi : 10.3406/quad.1987.2037

[5] Bernard Stiegler, états de choc : Bêtise et savoir au XXIe Siècle, Mille et une Nuit, 2012

[6] J. L. Moreno, The Future of Man's World,, New York Beacon House, Psychodrama Monographs, 1947

[7] Gilbert Simondon, L’individuqtion Psychique et Collective, à la lumière des notions de Forme, Information, Potentiel et Métastabilité, Paris, Editions Aubier, 1989 et 2007

[8]Satoshi Hamano, Architecutre no seitaikei: Johokankyo wa ikani sekkeisaretekitaka(The Ecology of Architecture), Chinese translation, Taiwan, 2011

Trust, Ethics and Legal Aspects of Social Computing

Andrew Power, Grainne Kirwan

Abstract. The development of a legal environment for virtual worlds presents issues of both law and ethics. The cross-border nature of online law and particularly law in virtual environments suggests that some lessons on its formation can be gained by looking at the development of international law, specifically the ideas of soft law, and adaptive governance. In assessing the ethical implications of such environments the network of online regulations, technical solutions and the privatization of legal remedies offer some direction. While legal systems in online virtual worlds require development, the ethical acceptability of actions in these worlds is somewhat clearer, and users need to take care to ensure that their behaviours do not harm others.

1 INTRODUCTION

Social networks and virtual worlds are becoming a more important and prevalent part of our real world with each passing month. Shirky [1] argues that the old view of online as a separate space, cyberspace, apart from the real world is fading. Now that computers and computer like smartphones have been so broadly adopted there is no separate cyberworld, just a more interconnected 'new' world. The internet augments real world social life rather than providing an alternative to it. Instead of becoming a separate cyberspace, our electronic networks are becoming embedded in real life [2]. The reason for this growth is in part down to the natural inclination of humans to want to form groups and interact with each other, combined with the increasing simplicity of the technology to allow it. As Shirky [2] states, "Communications tools don't get socially interesting until they get technologically boring. [The tool] has to have been around long enough that most of society is using it. It's when a technology becomes normal, then ubiquitous, and finally so pervasive as to be invisible, that the really profound changes happen."

Crime in a virtual world can take a number of forms. Some activities such as the theft of goods are relatively clear-cut whereas, private law issues such as harassment or commercial disputes are more complex. Online crime is defined as, crime committed using a computer and the internet to steal a person's identity or sell contraband or stalk victims or disrupt operations with malevolent programs. The IT security company Symantec [3] defines two categories of cybercrime, "Type I, examples of this type of cybercrime include but are not limited to phishing, theft or manipulation of data or services via hacking or viruses, identity theft, and bank or e-commerce fraud. Type II cybercrime includes, but is not limited to activities such as cyberstalking and harassment, child predation, extortion, blackmail, stock market manipulation, complex corporate espionage, and planning or carrying out terrorist activities". Types of crime can be categorized as internet enabled crimes, internet specific crimes and new crimes committed in a virtual world. The first two categories of online crime have been observed for many years and the third, which coincided with the growth in online virtual environments, is a more recent development. Internet enabled crimes are those crimes which existed offline but are facilitated by the Internet. These include credit card fraud, defamation, blackmail, obscenity, money laundering, and copyright infringement. Internet specific crimes are those that did not exist before the arrival of networked computing and more specifically the proliferation of the internet. These include, hacking, cyber vandalism, dissemination of viruses, denial of service attacks, and domain name hijacking. The third category of crimes committed in a virtual world arises when individuals are acting through their online avatars or alternate personas (the Sanskrit word avatara means incarnation). In computing an avatar is a representation of the user in the form of a three-dimensional model. Harassing another individual through their online representation may or may not be criminal but it is at the very least antisocial. It is also the case that that online activities can lead to very real crimes offline.

This paper aims to introduce some of the types of crimes which can occur in virtual worlds through a series of examples of actual virtual crimes, such as virtual sexual assault, theft, and child pornography. It should be noted that while the term 'crimes' will be used to describe these acts throughout the chapter, and the term 'criminals' assigned to the perpetrators, the actions are not necessarily criminal events under any offline legal system, and the perpetrators may not be considered criminal by a court of law. In some cases there have been offline consequences of the actions which are real criminal events, but in many cases no criminal prosecution is currently possible. Nevertheless, this is not to say that these virtual criminal behaviours are actually ethical, and the chapter also considers the impact of the behaviour on the individuals involved. Finally it is aimed to determine what the implications are for law formation in virtual worlds, along with an examination of how these should be implemented.

2 VIRTUAL WORLDS AND ONLINE CRIMES

Online theft of virtual goods has led to serious crimes offline. In 2008 a Russian member of the Platanium clan of an MMORPG (massively multiplayer online role-playing game) was assaulted in the Russian city of Ufa by a member of the rival Coo-clocks clan in retaliation for a virtual assault in a role playing game. The man died of his injuries en route to hospital [4]. Even if the activity does not spill over into the real world but remains online it is clear that crime can occur. In August 2005 a Japanese man was arrested for using software 'bots' to 'virtually' assault online characters in the computer game Lineage II and seal their virtual possessions. Bots, or web robots, are software applications that run automated tasks over the Internet. He was then able to sell these items through a Japanese auction website [5]. In October 2008, a Dutch court sentenced two teenagers to 360 hours of community service for 'virtually' beating up a classmate and stealing his digital goods [6]. In 2007 a Dutch teenager was arrested for stealing virtual furniture from 'rooms' in Habbo Hotel, a 3D social networking website; this virtual furniture was valued at €4,000 [7].

Internet child pornography is a topic which is eliciting greater attention from society and the media, as parents and caregivers become more aware of the risks to their children and law enforcement agencies become more aware of the techniques and strategies used by offenders. Sheldon and Howitt [8] indicate that at least in terms of convictions, internet child pornography is the major activity that constitutes Internet related sex crimes. An example of the kind of ethical controversies this subject can produce is the Wonderland area of Second Life which provided a place for role play of sexual activity with "child" avatars. This drew out many questions which are dealt with by Adams [9] and Kirwan and Power [10]. These include examining when the fantasy of illegality becomes illegal, the verification of participant's age, and the definition of harm in a virtual world. Online activity may be an outlet for harmful urges or an encouragement toward them; it may have a therapeutic role or alternatively promote the normalization of unacceptable behaviours.

In Britain a couple are divorcing after the wife discovered her husband's online alter-ego was having an affair online with another, virtual, woman [11]. This is interesting in that the "affair" was virtual and involved a relationship between the avatar of the husband and the avatar of another woman. Is it possible to be unfaithful to your real world partner by having your alter ego have an online only relationship? Clearly in the view of this man's wife it is and it hurt just as much, she said "His was the ultimate betrayal. He had been lying to me." Was this a question of trust, ethics, or just a lack of a shared understanding about the rules of a game vs. the rules of life?

3 ETHICS AND TRUST IN A VIRTUAL WORLD

Our view of what is ethical is informed by our world view and it is possible that more than one system of values can exist simultaneously. Isaiah Berlin [12] argued that when it comes to questions like "what is justice?" there is never a single answer. This leads to a variety of answers depending on the value systems in a given time and place. There can be no one value system that can accommodate all that is valuable. So there will be competing values systems even within the same community and at a given point in time. There is also no objective system to evaluate which is right and which is wrong (or less right!). Value systems are essential to the models through which we see ourselves and the world around us and they embody deeply held convictions. John Rawls [13,14] sought to develop a theory of justice suitable for governing political communities in the light of irreconcilable moral disagreements.

These debates are crucial in considering behaviour in online societies. Social networks will emerge in different ways and for different purposes and as such will require different value systems. Constructing systems of variable ethics and providing choice in online value systems will pose increasing challenges to states, individuals and systems of justice. To give one example, the behaviour considered correct and moral in an environment such as Grand Theft Auto will, one hopes, be quite different to that of Club Penguin. The world of Grand Theft Auto consists of a mixture of action, adventure, driving, and shooting and has gained controversy for its adult nature and violent themes. Club Penguin in contrast is aimed at young children who use cartoon penguins as avatars to play a series of games in a winter "polar" environment. Both in terms of the activities engaged in and the nature of the language used these environments could not be more different from an ethical perspective. However both conform to their own internal rule set for player behaviour.

This allows for the possibility of individual citizens being part not only of a number of different online societies with different standards of ethics, but that most or all of these may be different to the ethical standard assumed to be the norm when offline. This dichotomy or system of variable ethics may not have much societal impact if the online worlds are restricted to games, or infrequent visits to virtual worlds for entertainment. However as commercial interest, banks, and the state begin to move services online and explore virtual communities and service centres this issue becomes more prescient.

In opposition to the ideas of John Rawls mentioned earlier, Robert Nozick argued that the solution was not the reimagining of the state but its removal [15]. In his book 'Anarchy, State, and Utopia' Nozick makes the case for a minimal state limited to the most narrow of functions of protection of citizens against external force, theft and contract law. A state which moves beyond this narrow role will, he argues, lead to the violation of rights. The diminishing of the role of the state in the development of ethical standards, either by a Rawlian reimagining of the state or a Nozickian removal of the state for such matters, will in either case lead to a greater role for the individual in setting his or her own subjective ethical standard.

Online identities are not restricted by reality. They 'need not in any way correspond to a person's real life identity: people can make and remake themselves, choosing their gender and the details of their online presentation' [16]. Impression management is the process of controlling the impressions that other people form, and aspects of impression management normally outside our control in face-to-face interactions, can be controlled in online environments [17]. In the online context, we can easily manage and alter how other people see us in ways that were never before possible.

Given this reality can a personal attack against an avatar be construed as the equivalent of an attack against the person whom the avatar represents? The 'humanity' or otherwise of avatars in virtual worlds is important. Can they be considered equal to human victims of crimes? Has harm really been done? The answer to this lies both in the degree of separation the creator of the avatar has between their online and offline personas and their degree of attachment to their avatar. Spending a large amount of time 'in the skin' of our avatar can lead to strong feelings of association to the point where an attack on the avatar can feel like an attack on self. The degree to which a person experiences a strong sense of presence within a virtual world is discussed in detail by Kirwan [18]. It is also true that as we spend greater amounts of time online the differences between our online and offline personalities diminish. In part this is because it is just too much trouble to maintain two different personae but also because the distinction between the 'real' world and our online world are no longer meaningful. Shirky [2] outlines the problem of treating the internet as some sort of separate space or cyberspace when he states; "The internet augments real-world social life rather than providing an alternative to it. Instead of becoming a separate cyberspace, our electronic networks are becoming deeply embedded in real life". We only live in one world but an increasing portion of our time is spent interconnected to others though technology. It is not an alternative world it is just part of our new world.

Robert Putnam [19] wrote about the decline in social capital and described the declining vibrancy of American civil society, as evidenced by the reduced participation in community-based groups. His solution was in large part built on the 'development of networks, norms and social trust that facilitate coordination for mutual benefit' [20]. He considers that the pursuit of shared objectives provides a way for people to experience 'reciprocity' and thus helps to create webs of networks underpinned by shared values. The resulting high levels of social trust foster further cooperation between people and reduce the chances of anti-social conduct [21].

Rachel Botsman [22] makes the case that technology is enabling trust between strangers. Products like Swaptree and eBay which facilitate online trading only work in an environment of trust. Collaborative behaviours and trust mechanics are embedded in these systems. These networks mimic the ties that used to happen face-to-face but on a massive scale. Social networks and real-time technologies are taking us back to a system of bartering, trading and swapping where we have wired our world to share. This is happening in our neighbourhood, our schools, our workplaces, and on our Facebook network. This she calls collaborative consumption. We are moving from passive consumers, to creators, to active collaborators. This transition is actually a return to the behaviour we should be most comfortable with. As we are increasingly interconnected through social networks this is providing us with opportunities to express this social dimension and to be active in our many communities. Younger, citizens are developing networks of trust and confidence in virtual spaces which are informing their behaviour in their communities and informing their sense of the polis.

4 THE IMPACT ON VICTIMS OF VIRTUAL CRIME

There are a number of reactions that are evident in victims of crime, as outlined by Kirwan [18]. These vary according to both the type of crime and the coping strategy and personality of the individual victim, but can include Acute Stress Disorder (ASD) or Post-Traumatic Stress Disorder (PTSD), self-blaming for victimization, victim blaming (where others put all or partial blame for the victimization on the victim themselves), and a need for retribution. Virtual victimization, either of property crime or a crime against the person, should not be considered as severe as if a similar offence occurred in real life. However, it would be an error to believe that an online victimization has no effect on the victim at all.

Victim blaming appears to be particularly common for virtual crime. It has been argued that victims of virtual crime could easily escape. In Second Life, it is possible to engage in rape fantasies, where another player has control over the "victim's" avatar, but this is usually given with consent. There are suggestions that some individuals have been tricked into giving their consent, but even bearing this in mind, there has been widespread criticism by Second Life commentators of anyone who allows an attack to take place, as it is alleged that it is always possible to 'teleport' away from any situation, disconnect from the network connection or turn off their computer and thus end the event. It is clear that victims of virtual crime do seem to experience some victim blaming by others – they are in ways being blamed for not escaping their attacker. Those victims who experience the greatest degree of presence – those who are most immersed in the game - are probably those who are least likely to think of closing the application to escape. It should also be considered that a victim may experience discomfort at being victimized, even if they do escape relatively quickly. As in a real life crime, the initial stages of the attack may be confusing or upsetting enough to cause significant distress, even if the victim manages to escape quickly.

There is also some evidence of self-blaming by various victims of virtual crimes. Some victims refer to their relative naivety in the online world prior to victimization [23], and indicate that if they had been more experienced they may have realized what was happening sooner. There are also suggestions that a victim who is inexperienced with the virtual world's user interface may inadvertently give control of their avatar to another user. It is certain that empirical study needs to be completed on this topic before a definitive conclusion can be reached as to the degree of self-blaming which occurs.

There is also some evidence of limited symptoms of ASD in victims of virtual crimes, such as some anecdotal accounts of intrusive memories, emotional numbing and upset from victims of virtual sexual assault [24, 25]. While it is impossible to make an accurate judgment without a full psychological evaluation, it seems very unlikely that these victims would receive a clinical diagnosis of either ASD or PTSD. This is because there is no mention of either flashbacks or heightened autonomic arousal (possibly due to the lack of real danger to the victim's life). There are also several accounts of individuals who have experienced online victimization, but who do not see it as a serious assault and do not appear to experience any severe negative reaction. Those most at risk appear to be those who have previously experienced victimization of a real-life sexual assault, where the online attack has served to remind the victim of the previous attack. As such, while not a major risk, the possibility of developing ASD or PTSD is a factor that should be monitored in future victims of serious online assaults, especially those who have been previously victimized in real life.

Finally, there is substantial anecdotal evidence of a need for retribution in victims of virtual crimes. Similar reactions have been noted by other victims of crimes in virtual worlds, to the extent that in some cases victims have approached real world police forces seeking justice. This is possibly the strongest evidence that victims of virtual offences experience similar psychological reactions to victims of real life offences, although again, empirical evidence is lacking to date. As victims begin to seek justice, it seems necessary to consider the legal position of crimes in virtual worlds.

5 THE EVOLVING LAW ONLINE

Law online is inevitably international in nature given the cross border nature of the internet. As law making moved from the sole preserve of the state to supra state bodies such as the European Union and to entities such as the United Nations (UN), the International Monetary Fund (IMF), the World Bank, and the World Trade Organization (WTO), there was a move away from systems of command and control. As these changes occurred individual states had less autonomy, the importance of non-state actors grew and governance by peer review became important.

Another influence on the development of online law is the concept of soft law. Soft laws are those which consist of

informal rules which are non-binding but due to cultural norms or standards of conduct, have practical effect [26]. These are distinct from hard laws which are the rules and regulations that make up legal systems in the traditional sense. In the early days of the internet the instinct of governments was to solve the perceived problems of control by hard law. In the US the Clinton administration tried on many occasions to pass laws to control pornography online. The Communications Decency Act (CDA) was followed by the Child Online Protection Act (COPA) which was followed by the Children's Internet Protection Act (CHIPA). All were passed into law and all were challenged in the courts under freedom of speech issues.

Soft law offers techniques for compromise and cooperation between States and private actors. Soft law can provide opportunities for deliberation, systematic comparisons, and learning [27]. It may not commit a government to a policy but it may achieve the desired result by moral persuasion and peer pressure. It may also allow a state to engage with an issue otherwise impossible for domestic reasons and open the possibility for more substantive agreements in the future.

In considering the appropriate legal framework for the international realm of the internet the nature both of the activities taking place and the individuals and organizations using it need to be considered. The legitimacy or appropriateness of hard versus soft laws depends on the society they are seeking to legalize. In the context of online social networks soft laws have a power and potential for support which may make them more effective than the hard laws that might attempt to assert legitimacy. It is confluence of States, individuals, businesses, and other non-State actors that make up the legal, regulatory and technical web of behaviours that make the internet somewhat unique.

There are a number of views about the need for 'cyberlaws'. One is that rules for online activities in cyberspace need to come from territorial States [28]. The other is that there is a case for considering cyberspace as a different place where we can and should make new rules [29]. A third option is to look at the decentralization of law making, and the development of processes which do not seek to impose a framework of law but which allows one to emerge.

This could involve the creation of in-world systems of governance (controlled by software engineers, users, administrators, or a combination of these). Service providers would develop their own systems of governance and ethics. The law would come from the bottom up as users select the services, products and environment that match their own standards of behaviour and ethics. This would constitute a system of variable ethics. For example a user may choose to abide by the ethical norms in Grand Theft Auto and be quite comfortable with the notion of violent behaviour as a norm. Another user may be more comfortable in the ethical environment of Club Penguin. The ethical world is thus no longer normative but adaptable, variable or "fit for purpose". In this sense the ethical norms are not just variable but relative to the task at hand or the environment in which the citizen or user finds themselves. Relative ethics seems to be a contradiction in terms or perhaps indicative of a lack of moral clarity. This may be the view of some but an alternate view is that it moves the ethical framework by which a person lives their life away from a singularity such as church or state and towards the individuals own informed moral compass.

An approach suggested by Cannataci and Mifsud-Bonnici [30] is that 'there is developing a mesh of private and State rules and remedies which are independent and complementary'. The internet community can adopt rules and remedies based on their 'fitness for purpose'. State regulation may be appropriate to control certain activities, technical standards may be more appropriate in other situations, and private regulation may be appropriate where access to State courts or processes are impossible. Our understanding of justice may change as we see what emerges from un-coerced individual choice [31]. The appropriate legal or ethical framework on one context or virtual environment may be quite different in another.

Some aspects of what can and cannot be done, or even what may be considered right or wrong, will be determined by software engineers. They will find ways to prevent file sharing or illegal downloading or many other elements of our online activities. The blocking or filtering software that has largely removed the need for states to struggle with issues of censorship is being improved and refined all the time. This raises the question of the ethical landscape which results from coding. If the rules of the environment are set in part by programmers are we confident that the ethical norms of, for example, a young, male, college educated, Californian software engineer will necessarily match the needs or desires of all users? Private regulations also exist in the realm of codes of behaviour agreed amongst groups of users or laid down by commercial organizations that provide a service or social networking environment. The intertwining of State and private regulation is both inevitable and necessary to provide real-time solutions to millions of online customers and consumers.

Another part of the framework for considering law on the internet can be taken from the writing of Cooney and Lang [32]. They describe the recent development of learning-centred alternatives to traditional command-and-control regulatory frameworks, variously described as 'experimentalist' governance, 'reflexive' governance, or 'new' governance. Elements of these approaches contribute to what Cooney and Lang call adaptive governance. In this way all the sources of governance; user choice, code, private and state regulation, are all in constant flux as they both influence each other and improve and change overtime.

6 POLICING, PUNSHMENT & VICTIM SUPPORT

Online crimes with real world impact and risks should be under the remit of the traditional and appropriate enforcement agencies. This would include child pornography, online grooming of children, identity theft and appropriate hacking activities. However, in many cases the line is blurred, such as if a virtual attack is interpreted as an actual threat against the victim in real life. If an item is stolen in a virtual world, and the item can be judged to have an actual monetary value in real life, then it may also be possible to prosecute the thief in real life [33]. However, the line between a real life crime, and one which is purely virtual, is less coherent when the damages caused to the victim are emotional or psychological in nature, without any physical or monetary harm being caused. It is for these cases in particular that legal systems need to consider what the most appropriate course of action should be.

Policing of virtual worlds would most likely need to be unique to each world, if only because different worlds have differing social norms and definitions of acceptable and unacceptable behaviours. For example, players in an online war game such as Battlefield are unlikely to need a legal recourse if their avatar is killed when they lose, especially when the avatars come back to 'life' after a short time. However, if the same virtual murder occurred in an online world aimed at young children, it would obviously be much less acceptable. With this in mind, should it be obligatory for the creator of each virtual world to put in place a strict set of laws or regulations outlining what is and is not acceptable in the world, and ensuring that the virtual world is patrolled sufficiently well to ensure that all wrongdoings are observed and punished appropriately. An alternative is to make cybersocieties mirrors of the real world, where the police rely greatly on the citizens of the relevant society to report misconduct. On the other hand, this approach may also be open to abuse as one or more players could make unfounded allegations against another.

The punishment of virtual crime is often framed by a restorative justice approach. This refers to processes involving mediation between the offender and the victim [34]. Rather than focusing on the criminal activity itself, it focuses on the harm caused by the crime, and more specifically, the victims of the crime. It often involves a mediated meeting between the victim and the offender, where both are allowed to express sentiments and explanations, and the offender is given the opportunity to apologize. The aims of restorative justice are a satisfied victim, an offender who feels that they have been fairly dealt with, and reintegration of the community, rather than financial compensation or specific punishment. If the mediation does not meet the satisfaction of all involved, alternative punishments can then be considered. It would appear that the restorative justice approach is ideally suited for many virtual crimes as it allows the victim to feel that they have been heard, while allowing the community to remain cohesive. However, it should be noted that not all victims of real life crimes have felt satisfied by the process [35], and so in some online cases it may be inadequate or fail to satisfy those involved. It has been argued that virtual punishment is the appropriate recourse for crimes which occur in an online community [36]. In theft cases where the item has a 'real world' value, then it may be possible in some jurisdictions to enforce a 'real world' punishment also – perhaps a fine or a prison term.

Victims of real-life offences normally have relatively straightforward procedures available to them for the reporting of criminal offences. In online worlds, the reporting procedure is less clear, and the user may need to invest time and energy to determine how to report their experience. Although many online worlds have procedures for reporting misconduct, these are not always found to be satisfactory by victims if they wish to report more serious offences [23]. Similarly, reporting the occurrence to the administrators of the online world alone may not meet the victim's need for retribution, especially if they feel that they have experienced real-world harm because of the virtual crime. In those cases, the victim may prefer to approach the real-world authorities. To aid victims in this regard, many online worlds need to be clearer about their complaints procedures, and the possible outcomes of this. They may also need to be clearer about the possible repercussions of reporting virtual crimes to real world authorities.

Victims of real world crimes receive varying degrees of emotional, financial and legal aid, depending on the offence which occurred. In some cases, this aid is provided through charitable organizations, such as Victim Support, sometimes through government organizations, and also through informal supports such as family and friends. Financial aid is probably the least applicable to victims of virtual crime, as although theft of property can occur, it is unlikely to result in severe poverty for the victim. Also, because items with a designated real-world value are starting to be considered by real-world authorities, there is some possibility of financial recompense. Legal aid, both in terms of the provision of a lawyer and in terms of help in understanding the court system, can also be provided to real world victims. The legal situation is somewhat less clear for victims of virtual crimes, particularly where the punishment is meted out in the virtual world. But from the cases which have been publicized to date, it appears that the greatest need for assistance that online victims have is for emotional support. In some cases victims have sought this from other members of the online community, but the evidence of victim-blaming for virtual crimes which is apparent to date may result in increased upset for victims, instead of alleviating their distress.

7 CONCLUSIONS & FUTURE WORK

Cybersocieties have largely been making the rules up as they go, trying to deal with individual cases of virtual crime or anti-social behaviour, often without the action being criminalized in the community beforehand. In some cases this has been relatively successful, but in others victims of virtual offences appear to experience quite serious emotional reactions to their victimization, with limited acceptance of their reaction from others. With increasing numbers of both children and adults joining multiple online communities, it is important that adequate protection is provided to the cybercitizen.

These ideas of variable ethics (providing choice in online value system), soft law and adaptive governance offer lessons to the notion of a structure of laws for the internet. Systems of informal rules which may not be binding but have effect though a shared understanding of their benefits. Adaptable law which is flexible and open to change as knowledge develops. Agreements which include States and non-state actors, and which involve both the citizen and business. Soft law offers lessons on continuous learning in a changing environment, resulting in an evolving system of law and ethics and will pose increasing challenges to states, individuals and systems of justice.

Further work into the 'humanity' or otherwise of avatars in virtual worlds and the connection a user feels towards their avatar is important when considering the ethical response of users to each other. Further research also needs to be conducted in order to determine how widespread virtual crime actually is, and to establish how severely most victims react to it. The factors which lead to more severe reactions should then be identified. If virtual crime is determined to be a serious problem, with substantial effects on victims, then a greater focus needs to be placed on how online communities deal with this problem, and if legislation needs to be changed to reflect the psychological and emotional consequences of victimization. It should also be established if there are distinct or unique motives for online crime which do not apply to offline crime and how can these be combated.

REFERENCES

[1] C. Shirky, *Cognitive Surplus*, Penguin, London, (2010).

[2] C. Shirky, *Here comes everybody*, Penguin, London, (2009).

[3] Symantec (n.d.) *What is Cybercrime?* http://www.symantec.com/norton/cybercrime/definition.jsp (2012).

[4] F. Truta, *Russia - Gamer Kills Gamer over Gamer Killing Gamer... Er, In-Game!* http://news.softpedia.com/news/Russia-Gamer-Kills-Gamer-over-Gamer-Killing-Gamer-Er-In-Game-76619.shtml (2008).

[5] W. Knight, *Computer characters mugged in virtual crime spree*, http://www.newscientist.com/article/dn7865, (2005).

[6] D. McNeill, Virtual killer faces real jail after murder by mouse, *The Independent*, http://www.independent.co.uk/life-style/gadgets-and-tech/news/virtual-killer-faces-real-jail-after-murder-by-mouse-972680.html, (2008).

[7] BBC, Virtual theft leads to arrest, http://news.bbc.co.uk/2/hi/technology/7094764.stm, (2007).

[8] K. Sheldon and D. Howitt, *Sex Offenders and the Internet*, Wiley, Chichester, (2007).

[9] A.A. Adams, Virtual Sex with Child Avatars (pp. 55-72). In: *Emerging Ethical Issues of Life in Virtual Worlds*. C. Wankel and S. Malleck (eds.) Information Age Publishing, Charlotte, North Carolina (2010).

[10] G. Kirwan and A. Power, *The Psychology of Cybercrime: Concepts and Principles*, Information Science Reference, Hershey, PA (2011).

[11] S. Morris, Second Life affair leads to couple's real-life divorce, *Guardian*, http://www.guardian.co.uk/technology/2008/nov/14/second-life-virtual-worlds-divorce (2008)

[12] I. Berlin, *Concepts and Categories: Philosophical Essays*, Oxford University Press, Oxford, (1980).

[13] J. Rawls, *A Theory of Justice*, Oxford University Press, Oxford, (1973).

[14] J. Rawls, *Political Liberalism*, Columbia University Press, New York, (1996).

[15] R. Nozick, *Anarchy, State, and Utopia*, Basic Books, Inc., UK, (1974)

[16] J. Mnookin, Virtual(ly) Law: The Emergence of Law in LambdaMOO. *Journal of Computer-Mediated Communication*. 2(1), http://www.ascusc.org/jcmc/vol2/issue1/lambda.html (1996).

[17] A. Chester, & D. Bretherton, Impression management and identity online. In A. Joinson, K. McKenna, T. Postmes, & U. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 223-236), Oxford University Press, New York, (2007).

[18] G. Kirwan, Presence and the Victims of Crime in Online Virtual Worlds. *Proceedings of Presence 2009* – the 12th Annual International Workshop on Presence, International Society for Presence Research, November 11-13, Los Angeles, California. http://astro.temple.edu/~tuc16417/papers/Kirwan.pdf, (2009).

[19] R. Putnam, 'Bowling Alone: America's Declining Social Capital', *Journal of Democracy*, 6 (1): 65–78, (1995).

[20] F. Locke, P. Rowe, and R. Oliver, *The Impact of Participation in the Community Service Component of the Student Work and Service Program (SWASP) on student's continuing Involvement in the Voluntary, Community-Based Sector*, http://www.envision.ca/pdf/cscpub/SwaspResearchPaper2004.pdf, (2004).

[21] B. Hoskins, *A framework for the creation of indicators on active citizenship and education and training for active citizenship*, Ispra, Joint Research Centre, (2006).

[22] R. Botsman and R. Rogers, *What's Mine Is Yours: How Collaborative Consumption is Changing the Way We Live*, Collins, New York, (2011).

[23] E. Jay, *Rape in Cyberspace*, https://lists.secondlife.com/pipermail/educators/2007-May/009237.html, (2007).

[24] J. Dibbell, *A Rape in Cyberpspace*, http://loki.stockton.edu/~kinsellt/stuff/dibbelrapeincyberspace.html, (1993).

[25] J. Dibbell, *A Rape in Cyberspace*, http://www.juliandibbell.com/texts/bungle.html, (1998).

[26] P. Burgess, What's So European About the European Union?: Legitimacy Between Institution and Identity. *European Journal of Social Theory*, (5), 467, (2002).

[27] A. Schäfer, Resolving Deadlock: Why International Organizations Introduce Soft Law, *European Law Journal*, (12)2, 194-208, Blackwell Publishing Ltd., Oxford, (2006).

[28] J.L. Goldsmith, Against cyberanarchy, *University of Chicago Law Review*, (65), 1199, (1998).

[29] D. Johnson and D. Post, Law and Boarders – The Rise of Law in Cyberspace, *The Stanford Law Review*, (48)5, 1367 – 1402, (1996).

[30] J. Cannataci and P. Mifsud-Bonnici, Weaving the Mesh: Finding Remedies in Cyberspace, *International Review of Law, Computers & Technology*, (21)1, 59-78, (2007).

[31] D. Post, Governing Cyberspace, *The Wayne Law Review*, Vol.43, No.1 155- 171, (1996).

[32] R. Cooney and A. Lang, Taking Uncertainty Seriously: Adaptive Governance and International Trade, *European Journal of International Law*, (18), 523, (2007).

[33] R. Hof, *Real Threat to Virtual Goods in Second Life*, http://www.businessweek.com/the_thread/techbeat/archives/2006/11/real_threat_to.html, (2006).

[34] D. Howitt, *Introduction to Forensic and Criminal Psychology* (3rd edition). Pearson, (2009).

[35] J.A. Wemmers and K. Cyr, Victims' perspectives on restorative justice: how much involvement are victims looking for? *International Review of Victimology*, (11), 259-274, (2006).

[36] R.C. McKinnon, Punishing the persona: Correctional Strategies for the Virtual Offender. In S. Jones (ed.) *The Undernet: The Internet and the Other*. Sage (1997).

Facebook's user: product of the network or 'craft consumer'?

Ekaterina Netchitailova[1]

Abstract. There is an ongoing debate about the role of the users of Facebook within the network. On the one hand, the user of Facebook can be seen as a 'product' of the network and a free labour force working for Facebook for free, but on the other hand, the same user can be seen as a 'craft consumer', participating in the 'trickery' within the network as well as taking part in making policy of Facebook, as the failed initiative of Beacon demonstrates. The role of the user within the network is usually analysed either by using critical Internet Theory (critical studies of communication, as advanced by Fuchs, 2008, 2010, 2011) where the user emerges as a 'prosumer commodity', a commodity which is produced, sold and consumed, or through 'celebratory media studies', where the user is seen as an active agent who takes an active role in making Facebook. Both these approaches tend to be either very optimistic or pessimistic in looking at the role of the user within such a network as Facebook. However, a new approach is needed which encompasses both views. We propose in this paper to go back to the notion of a 'craft consumer' as proposed by Cambell (2005) [1] where the user is crafting things he consumes, including Facebook's usage.

1 INTRODUCTION

Facebook is many different things: it is a useful tool to stay in touch, a platform for organising groups and petitions, a means to portray oneself in an 'interesting' way and Facebook is also ultimately a corporation, and whose main drive is profit.

The way we choose to look at Facebook determines the way we analyse the role of the user of Facebook. Take Facebook and its greeting which says 'Facebook helps you to connect and share with the people in your life', and Facebook emerges indeed as a wonderful tool, which helps us to find lost classmates, stay in touch with friends and organise all kinds of events. Here Facebook emerges as a Web 2.0 tool, where users are not only consumers of the content but also are its creators.

However, if we look at Facebook as a corporation, another picture can be drawn. Facebook is ultimately a capitalistic structure, pursuing profit and with a dubious privacy policy. As the privacy policy of Facebook says: "For content that is covered by intellectual property rights, like photos and videos ('IP content') you specifically give us the following permission, subject to your privacy and application settings: you grant us a non-exclusive permission, subject to your privacy and application settings: you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post on or in connection with Facebook ('IP licence'). This IP licence ends when you delete your IP content or your account unless your content has been shared with others, and they have not deleted it." (www.facebook.com). [2]

We can also find the following paragraph:

"When you access Facebook from a computer, mobile phone, or other device, we may collect information from that device about your browser type, location, and IP address, as well as the pages you visit." (www.facebook.com)

This means that Facebook collects information about us. It can also sell information about us to advertisers, and here the user emerges as someone who is used and actually works for free for Facebook.

These two views on Facebook are reflected in the current research on Facebook. On the one hand we have what can be called 'celebratory cultural studies' (Fuchs, 2011) [3], led by such researchers as boyd (2008,2010) [4] and Jenkins (2006) [5] and which view online social networks as spaces for community-building, friendship formation and autonomous spaces where people can have 'fun' and take an active part in network's creation. Here the user is seen as an active agent who participates in the art of making of everyday life, including his involvement with Facebook. On the other hand, however, we have critical studies of communication, led by Fuchs (2008, 2010, 2011) which see online social networks as sites of domination and oppression, where user is used for the purposes of the corporation.

Both of these views do not interact with each other in the current analysis of online social network and as a result an important part of the analysis is missing. By focussing only on the user we miss the societal aspects of the network, its macro-context and how it is shaped by capitalism. But by focussing only on the oppressive side of an online social network, we miss the perspective of the user and the concept of 'joy' and 'playfulness' within the network. As Dwayne Winseck argues in his discussion with Christian Fuchs (2011), by reducing media and communication to instruments of domination there is a danger to overlook the links between communication and media and pleasure and joy.

We think that in the analysis of such a network as Facebook, it is important to look at both how the user is 'exploited' by Facebook, by underlying the capitalistic structure of Facebook but also at how the user makes Facebook 'his own', reworks it and has fun with it. We propose to look at Facebook's user as a 'craft consumer', who not only consumes the content on Facebook but also participates in making 'craft' out of it.

2 Facebook as Web 2.0

Facebook can be seen as a part of Web 2.0/Web 3.0 where users are not only consumers of the content but also are its creators.

[1] Dept. of Sociology, Sheffield Hallam University, S1 1WB, UK. Email: ekaterina.p.netchitailova@student.shu.ac.uk

In the first phase of the development of the Internet, World Wide Web was dominated by hyperlinked textual structures, called Web 1.0. It is characterized by text-based sites and is mostly a system of cognition. (Fuchs, 2008) [6] However, with the rise of such sites as Youtube, MySpace and Facebook, both communication and cooperation became important features of the Web. The Web characterized by communication is called Web 2.0. Web 3.0, on the other hand, is not only communicative but also cooperative. An example of Web 3.0 is Wikipedia, where everyone can participate in the creation of the content. Thus, Fuchs says that Web 1.0 (where we mostly read the text but do not participate) is a tool for thought, Web 2.0 is a medium for human communication and Web 3.0 technologies "are networked digital technologies that support human cooperation." (Fuchs, 2008, p. 127)

The main thought associated with Web 2.0 platforms is that people take a more pro-active approach in their creation.

Jenkins in his 'Convergence Culture' (2006) talks about three new trends which have been shaping media lately. These are media convergence, participatory culture and collective intelligence.

By media convergence he means that today the content flows across multiple media platforms, different media industries cooperate with one another and media audiences have a greater choice about where to seek content. An example of media convergence on Facebook would be many posts of users where they provide links to different sites, including Youtube or CNN. This permits the user to get different kind of news and information and raises awareness about issues which otherwise would have remained unknown.

An example of media convergence would be Obama's presidential campaign in 2008.

The use of different media outlets and especially of online social networks was central to the election win. Obama used Twitter and Facebook, blogs and video-sharing sites including YouTube, to spread his political views and rally supporters. Staff of Obama directly responded to voters' questions about Obama's policies and views via social networking sites. As Ranjit Mathoda wrote on his blog: "…Senator Barack Obama understood that you could use the Web to lower the cost of building a political brand, create a sense of connection and engagement, and dispense with the command and control method of governing to allow people to self-organize to do the work." (from www.mathoda.com) [7]

In April 2010 President Obama announced that he was seeking re-election to the highest office via YouTube video.

By participatory culture Jenkins means that people today are actively participating in the creation of media content.

> "Rather than talking about media producers and consumers as occupying separate roles, we might now see them as participants who interact with each other according to a new set of rules that none of fully understands." (Jenkins, 2006, p. 3)

And by collective intelligence Jenkins means that the consumption of media has become a collective process, where producers and consumers of media work side by side.

> "Convergence requires media companies to rethink old assumptions about what it means to consume media, assumptions that shape both programming and marketing decisions. If old consumers were assumed to be passive, the new consumers are active. If old consumers were predictable and stayed where you told them to stay, then new consumers are migratory, showing a declining loyalty to networks or media. If old consumers were isolated individuals, the new consumers are more socially connected. If the work of media consumers was once silent and invisible, the new consumers are now noisy and public." (Jenkins H., 2006, p. 19)

Jenkins gives an example of the reality show 'Survivor' whose viewers created an online forum, serving as an important platform for discussing the show, but also on some instances as a catalyst of changes in the show itself and as an important exchange of learning between viewers on different issues, not necessary limited to the show.

Thus, according to Jenkins, despite the increasing influence of big corporations, consumers and audiences can still play an active role in the cultural formation.

The example of active audience on Facebook can be seen in the reaction of its users to some of the initiatives taken by Facebook's owners.

On November 6, 2007 Facebook launched Beacon, a controversial social advertising system, that sent data from external websites to Facebook, allegedly in order to allow targeted advertisements and so that users could share activities with their friends.

However, as soon as it was launched it created considerable controversy, due to privacy concerns. People did not want the information about their purchases on the Internet to appear on Facebook's news feed for everyone to see. There was a story about a guy who had bought an engagement ring for his girlfriend, planned as a surprise, but this news appeared on Facebook for everyone to see. As this person complained:

> "I purchased a diamond engagement ring set from overstock in preparation for a New Year's surprise for my girlfriend. Please note that this was something meant to be very special, and also very private at this point (for obvious reasons). Within hours, I received a shocking call from one of my best friends of surprise and "congratulations" for getting engaged.(!!!)
>
> Imagine my horror when I learned that overstock had published the details of my purchase (including a link to the item and its price) on my public Facebook news feed, as well as notifications to all of my friends. ALL OF MY FRIENDS, including my girlfriend, and all of her friends, etc..."
>
> (from http://forrester.typepad.com/groundswell/2007/11/close-encounter.html) [8]

That same month a civic action group MoveOn.org created a Facebook group and online petition asking Facebook not to publish users' activity from other websites without explicit permission from a user. In ten days the group had 50,000

members. Facebook changed Beacon so that users had first to approve any information from external websites appearing on their news feed. However, it was found that the information from external websites was still collected by Facebook which provoked further controversy and angry reactions from Facebook's users.
In response Facebook announced in December that people could opt out of Beacon and Mark Zuckerberg apologized to Facebook's users.

As Scott Karp remarks in his article 'Facebook Beacon: A Cautionary Tale About New Media Monopolies' (2007) [9] the whole story with Beacon is much more interesting and important to the evolution of media than simply the reason why Beacon did not work.

Previously media companies could have complete control over their content. Even if we do not like advertisements on TV, we still watch the TV. Media companies have complete control over a TV channel, where a consumer has a little choice. However, with the advance of the Internet, the user has also a control over the content. The nature of monopoly has changed. Facebook is not really a monopoly, it simply has high switching costs.

> "So Facebook got caught in the perfect storm of believing it had a monopoly - when it didn't - and having the unprecedented technical capacity to abuse the privilege that it didn't actually have…It may well be that natural monopolies in media which drove the media business for the last century - are dead. And without monopoly control, you don't have license to exploit your audience, i.e. your users." (Scott Karp, 2007, from http://www.dmwmedia.com/news/2007/12/03/facebook-beacon:-cautionary-tale-about-new-media-monopolies, retrieved on 12.02.2011)

Beacon initiative showed that Facebook users want to have a say in how Facebook was run.

3 Facebook as corporation

While the initiative with Beacon was successfully sabotaged by Facebook's users, the participation of Facebook's users in how the site is run is not a straightforward one. When, in 2010, Facebook changed its privacy settings, many users started to complain, but the network effectively ignored the complaints and maintained the changes. This shows that Facebook as corporation makes the final decision about how it is run and its privacy policy clearly shows that data of users is used for advertisements purposes. Information on Facebook posted by its users provides invaluable knowledge to many corporations (including Facebook itself) and companies. Thrift (2005) [10] talks about knowledge economy, which underlines the current capitalistic society, where "knowledges that are transmitted through gossip and small talk which often prove surprisingly important are able to be captured and made into opportunities for profit." (Beer, 2008, p. 523) [11]

On Facebook we engage constantly with gossip and small talk and this can be used by many companies to target their advertisements.

And this leads to the following question. Are we indeed customers of Facebook or are we simply its product, as Andrew Brown asks rightly in his article "Facebook is not your friend." [12]

> "Anyone who supposes that Facebook's users are its customer has got the business model precisely backwards. Users pay nothing, because we aren't customers, but product. The customers are the advertisers to whom Facebook sells the information users hand over, knowingly or not. " (Brown A., 2010, http://www.guardian.co.uk/commentisfree/andrewbrown/2010/may/14/facebook-not-your-friend)

Even games and quizzes can be regarded as another tool to collect more information about us. Almost everything on Facebook is a means to harvest data about its users and therefore, Facebook is much more complicated than a wonderful tool to stay in touch with people. It is also a powerful advertising machine, a sophisticated business model, and the exchange on Facebook is two-sided. We get a tool to communicate with our friends, while in exchange we provide information about ourselves, which can be used by the government, advertising agencies, market research companies and Facebook itself.

Alvin Toffler (1980) coined the term *prosumer* within information society. Axel Bruns (2007) [13] applied this term to new media and coined the term produsers - where users become producers of digital knowledge and technology.

> "Produsage, then, can be roughly defined as a mode of collaborative content creation which is led by users or at least crucially involves users as producers - where, in other words, the user acts as a hybrid user/producer, virtually throughout the production process." (Bruns, 2007, p 3)

As Trebor Scholz (2010) [14] argues, we produce economic value for Facebook mainly in three ways: 1. providing information for advertisers, 2. providing unpaid services and volunteer work, and 3. providing numerous data for researchers and marketers.

Providing unpaid services and volunteer work is especially interesting, as Facebook basically users the labour of Facebook users for free. Scholz mentions that many Facebook users provide willingly their time and energy for Facebook use. The example is the translation application, where users translate Facebook into different languages totally for free. Roughly ten thousand people participated in the application which allowed the Facebook to be read and used in many languages, besides English.

As Fuchs says:

> "If users become productive, then in terms of Marxian class theory this means that they also produce surplus value and are exploited by capital as for Marx productive labour is labour generating surplus. Therefore the exploitation of surplus value in cases like Google, YouTube, MySpace, or Facebook is not merely accomplished by those who are employed by these corporations for programming, updating, and maintaining the soft- and hardware, performing marketing activities, and so on, but by wage labour and produsers who

> engage in the production of user-generated content." (Fuchs, Ch., 2009, p. 30) [15]

Users of Facebook also provide data and content for the site, making it more appealing for use, through photos, comments, etc. One of the strategies employed by such corporations as Facebook is to lure the users through the promise of free service, who in turn produce content. This content, in turn, is sold to third-party advertisers.

Maurizo Lazzarato introduced the term 'immaterial labour', which means "labour that produces the informational and cultural content of the commodity." (Lazzarato M., 1996, p. 133) [16] This term was popularized by Michael Hardt and Antonio Negri who said that immaterial labour is labour "that creates immaterial products, such as knowledge, information, communication, a relationship, or an emotional response." (in Fuchs Ch., 2011, p. 299) [17] For them the main purpose of immaterial labour is to create communication, social relations and cooperation. Knowledge produced by this way would be exploited by capital. "The common (…) has become the locus of surplus value. Exploitation is the private appropriation of part or all of the value that has been produced as common." (in Fuchs Ch., 2011, p. 299) [18]

As Fuchs explains the Internet is part for the commons because all humans need to communicate in order to exist. But, as he continues, "the actual reality of the Internet is that large parts of it are controlled by corporations and 'immaterial' online labour is exploited and turned into surplus value in the form of the advertising-based Internet prosumer commodity." (Fuchs, 2011, p. 299) [19]

Fuchs actually prefers the term 'knowledge labour' since 'immaterial labour' might mean that there are two substances of the world - matter and mind.

Knowledge labour is the labour that works for free in the Internet economy.

> "The concept of free labour has gained particular importance with the rise of web 2.0 in which capital is accumulated by providing free access. Accumulation here is dependent on the number of users and the content they provide. They are not paid for the content, but the more content and the more users join the more profit can be made by advertisements. Hence the users are exploited - they produce digital content for free in non-wage labour relationship." (Fuchs, 2011, p. 299) [20]

Capitalism's imperative is to accumulate more capital. In order to achieve this, capitalists either have to prolong the working day (then it is called absolute value production) or to increase the productivity of labour (relative surplus value production). (Fuchs, 2011) In the case of relative surplus value production productivity is increased so that more commodities and more surplus value are produced in the same period as previously.

Targeted Internet advertising can be called relative surplus value production. The advertisements are produced by advertising company's wage workers but also by users of the online social networks, whose content in the profiles and transaction data is used to make advertisements. Users also produce content for free for Facebook itself, and thus, provide unpaid labour, which Fuchs terms also 'play-labour'. (Fuchs, 2011). Users use such sites as entertainment mainly and usually in their free time. But without realizing it, in their free time they actually continue working for free for numerous Internet sites, by posting comments, updating profiles and by buying and selling things.

However, our argument is that the relationship between Facebook and its users is more complicated than seeing Facebook as 'exploiting' its users. Most users to whom I talked do not mind that Facebook sells their data to advertisers, provided it treats them with respect and does not intervene with their activities on the network. Moreover, numerous examples of 'trickery' and 'détournement' on Facebook can be seen as a response of users to Facebook's policy and as a demonstration that users of Facebook do not embrace Facebook without thinking but reflect about what it means and what Facebook represents.

4 'Trickery' on Facebook

Vejby and Wittkower in "Facebook and Philosophy" (2010) [21] talk about how users approach actively the culture around us through what they call 'détournement', which "refers to the subversion of pre-existing artistic productions by altering them, giving them a new meaning and placing them with a new context." (Vejby &Wittkower, 2010, p. 104)

They give an example of how users reacted to the privacy changes announced by Facebook by approaching changes ironically and through a play of words. They quoted also my status update in their chapter:

> **"Ekaterina Netchitailova** if you don't know, as of today, Facebook will automatically index all your info on Google, which allows everyone to view it. To change this option, go to Settings - -> Privacy Settings --> Search - -> then UN-CLICK the box that says 'Allow indexing'. Facebook kept this one quiet. Copy and paste onto your status for all your friends ASAP." (Wittkower, 2010, p. 105)

After this status update another one follows from a different user:

> "**David Graf** If you don't know, as of today, Facebook will automatically start plunging the Earth into the Sun. To change this option, go to Settings - -> Planetary Settings - -> Trajectory then UN-CLICK the box that says 'Apocalypse'. Facebook kept this one quiet. Copy and paste onto your status for all to see." (Wittkower D, 2010, p. 105)

And shortly afterwards another update appears:

> "**Dale Miller** If you don't, as of today, Facebook staff will be allowed to eat your children and pets. To turn this option off, go to Settings - -> Privacy Settings - -> then Meals. Click the top two boxes to prevent the employees of Facebook from eating your beloved children and pets. Copy this to your status to warn your friends." (Wittkower, 2010, p. 105)

One of my friends posted the following status update:

"WARNING: New privacy issue with Facebook! As of tomorrow, Facebook will creep into your bathroom when you're in the shower, smack your arse, and then steal your clothes and towel. To change this option, go to Privacy Settings > Personal Settings > Bathroom Settings > Smacking and Stealing Settings, and uncheck the Shenanigans box. Facebook kept this one quiet. Copy and paste on your status to alert the unaware"

This playful interchange allows Facebook's users to actively react to Facebook's policy and approach media content as active agents.

> "This kind of play may be silly, but it is significant. Of course, we should be concerned about privacy and Google-indexing of our Facebook posts, but the sense of participation and playful ridicule helps us to approach the media and culture around as active agents rather than passive recipients. It may not be the fullest from of political agency, but it's an indication of the kind of active irony which online culture is absolutely full of, and represents a kind of resistance and subversion." (Vejby& Wittkower, 2010, p. 105-106) [22]

There are many other examples of détournement on Facebook which demonstrate that users (at least some) think about Facebook and make 'fun' of it. One example is a group which is dedicated to art and has a special photo folder with references to Facebook as a part of culture and everyday life.
For instance, there is one picture which says:

"Do you want to make money from Facebook? It's easy. Just go to your Account settings, deactivate your account and go to Work!"

Another picture makes fun of the relationship status of Facebook. The text on the picture, on which a man and a woman lie in bed, shows their discussion in the following way: The woman says: "So? Is this it? Are we a couple now?, the man replies: "I don't know...I like this...I just...I don't know..." to which the woman says: "Well...Will you be my 'It's complicated on Facebook?'

And there is another picture which shows a woman in front of the computer with a text which says: "Now I have 3250 friends...I can share with them my solitude."

These instances of the playful use of Facebook might appear as silly, but they have an important point. They show that people, in their own way, not only make fun of Facebook but also reflect on the issues related to Facebook: its association with a waste of time, its influence on how we view friendships and community, and the fact that any activity on Facebook (like a status update or a new relationship status) is taken seriously by our Facebook 'friends'.

This détournement is actually an example of 'excorporation' discussed by John Fiske (1989) [23]. For him excorporation is "the process by which the subordinate make their own culture out of the resources and commodities provided by the dominant system, and this is central to popular culture, for in an industrial society, the only resources from which the subordinate can make their own subcultures are those provided by the system that subordinates them. There is no 'authentic' folk culture to provide an alternative, and so popular culture is necessary the art of making do with what is available. This means that the study of popular culture requires the study not only of the cultural commodities out of which it is made, but also of the ways that people use them. The latter are far more creative and varied than the former." (Fiske, 1989, p. 15)

Fiske gives an example of the commodity of jeans. Jeans are a perfect product of capitalism, many brands compete with each other to sell it to people and jeans are one of the most wearable item. But there are ways in which people, while still wearing them, manage to give an oppositional meaning to jeans, by 'debranding' them -by tie-dying them, bleaching irregularly or wearing them in a particular way. Another example that he gives is that of advertisements. We are constantly bombarded by advertisements from all corners in late capitalism, but people manage to turn advertisements into popular art, by playing with them and reworking them. For instance, children in Australia changed a 1982 beer commercial into a playground rhyme by singing: "How do you feel when you're are having a fuck, under a truck, and the truck rolls off? I feel like a Tooheys, I feel like a Tooheys, I feel like a Tooheys or two." (Fiske, 1989, p. 31)

Fiske reminds us of the 'trickery' term used by de Certeau, which is at the heart of popular culture:

> "The actual order of things is precisely what 'popular' tactics turn to their own ends, without any illusion that it will change any time soon. Though elsewhere it is exploited by a dominant power or simply denied by an ideological discourse, here order is *tricked* by an art. Into the institution to be served are thus insinuated styles of social exchange, technical inventions and moral resistance, that is, an economy of the 'gift' (generosities, for which one expects a return), an aesthetics of 'tricks' (artists' operations) and an ethics of *tenacity* (countless ways of refusing to accord the established order the status of a law, a meaning or a fatality." (in Fiske, 1989, p. 38)

The examples of playful interpretation of Facebook, like for instance, a picture which says: "I once had a life...when some idiot came and told me to make a Facebook account" or a text which says: "Spending a day on Facebook has once again fooled me into believing I have an actual social life" can be seen as an example of such excorporation or trickery on Facebook, as well numerous groups which actually discuss Facebook as corporation and compare it to Panopticon. These examples demonstrate that "the creativity of popular culture lies not in the production of commodities so much as in the productive use of industrial commodities. The art of people is the art of 'making do'. The culture of everyday life lies in the creative, discriminating use of the resources that capitalism provides." (Fiske, 1989, p. 28)

The user of Facebook then emerges as not only as a commodity, working for free for Facebook, but as a 'craft consumer' (Beer 2010, Cambell, 2005) [24], a consumer as defined by Colin Cambell, who has an active approach to the culture around him and participates in its creation. The definition proposed by Cambell "rejects any suggestion that the contemporary consumer is simply the helpless puppet of external forces." (Cambell, 2005. p. 24) [25] but an active agent involved in choosing the culture around him in a creative way. Then the power within Facebook is not only the power of Facebook as a corporation and

the power of groups of individuals to create groups to oppose the regime and status-quo, but also the power to be creative.

Building profiles (while according to some categories as defined by Facebook) is then a creative and in a way a powerful act. Putting status updates and talking with friends is an act of freedom, freedom to conduct one's everyday life as one sees fit.

5 Conclusion

The relationship between Facebook and its users is not a straightforward one. On the one hand, the user of Facebook can be seen as its product working for free for corporation, but, on the other hand, the same user can be seen as a 'craft consumer' actively engaging with the content of the network and 'having fun' with it.

So far, most studies either focus on the positive aspects of the network or the negative ones. However, a new direction is needed where critical theory of communication and media studies would incorporate popular culture for the analysis of such networks as Facebook in this society.

REFERENCES

[1] C. Cambell, 'The Craft Consumer: Culture, Craft and Consumption in a Postmodern Society', *Journal of Consumer Culture* 5 (1): 23-42 (2005)
[2] www.facebook.com
[3] Ch. Fuchs & D. Winseck, 'Critical Media and Communication Studies Today. A Conversation,' *TripleC* 9(2): 247-271, (2011)
[4]D. boyd, Taken out of context: American Teen Sociality in Networked Publics. PhD thesis. (2010)
[5] H. Jenkins, *Convergence Culture: where old and new media collide*, New York, University Press (2006)
[6] Ch. Fuchs, *Internet and Society. Social Theory in the Information Age*. Routledge. New York, London, (2008)
[7] www.mathoda.com
[8] http://forrester.typepad.com/groundswell/2007/11/close-encounter.html
[9] S. Karp, Facebook Beacon: A Cautionary Tale About New Media Monopolies. In www.dmwmedia.com, December 3, 2007.
[10] N. Thrift, *Knowing Capitalism*, Sage Publications. (2005)
[11] D. Beer, 'Social network(ing) sites…revisiting the story so far: A response to danah boyd & Nicole Ellison', *Journal of Computer-Mediated Communication*, 13, pp. 516-529. (2008)
[12] A. Brown, 'Facebook is not your friend', in http://www.guardian.co.uk/commentisfree/andrewbrown/2010/may/14/facebook-not-your-friend) (2010)
[13] A. Bruns, Produsage, generation C, and their effects on the democratic
process. In *Proceedings of the conference: Media in transition 5, MIT, Boston*.
http://web.mit.edu/comm-forum/mit5/papers/Bruns.pdf, (2007)
[14] T. Scholz, Facebook as Playground and Factory, in *Facebook and Philosophy*, D. Wittkower (ed.), Carus Publishing Company, (2010)
[15] Ch. Fuchs, 'Web 2.0, Prosumption, and Surveillance', *Surveillance & Society* 8(3), 288-309, (2009)
[16] M. Lazzarato, Immaterial Labour. In *Radical thought in Italy*, Virno, P and Hardt, M (eds), 133-146, Minneapolis, MN: University of Minnesota Press
[17] Ch. Fuchs, 'Web 2.0, Prosumption, and Surveillance', *Surveillance & Society* 8(3), 288-309, (2011)
[18] see [17]
[19] see [17]
[20] see [17]
[21] R. Vejby & D. Wittkower, Spectacle 2.0?, in D. Wittkower, (ed) *Facebook and Philosophy*. Carus Publishing Company, (2010)
[22] see [21]
[23] J. Fiske, *Understanding Popular Culture*, Routledge, (1989)
[24] D. Beer, 'Consumption, Prosumption and Participatory Web Cultures: An introduction', *Journal of Consumer Culture*, 10:3, (2010)
[25] C. Cambell, The Craft Consumer: Culture, Craft and Consumption in a postmodern Society. In *Journal of Consumer Culture*, vol. 5, no. 1, 23-42. 920050

Resorts behind the Construction of the Expositional Self on Facebook

Greti Iulia Ivana[1]

Abstract. The concept of self presentation, as developed by Goffman, has had a decisive influence on the literature about social networking sites. In the current paper, I explore some implications of what Hogan describes as a shift from presentation to exposure of the self, a phenomenon which is specific to the online environment. Drawing from Bourdieu and Baudrillard, I argue that the consumption practices, or more broadly, the lifestyle that the users expose through Facebook are a tool for the objectification and promotion of the self to a specific reference group.

Keywords: self, presentation, exhibition, objectification, Goffman, Bourdieu.

1 INTRODUCTION

The fast development of social network websites in the last 5 years has drawn the attention of researchers trying to explain their success and explore their implications. By far the fastest expanding such site is Facebook, counting an impressive 600 million active members in January 2011. Some of the key features that I believe individualize this network are the custom of presenting information that makes the user identifiable (such as real name and eloquent pictures) and the general tendency of creating a social network that comprises mainly of people with whom the user has had face to face interactions. Given the atypical amount of self disclosure as compared to most of the online environment and the strong link with the offline social universe, Facebook has been analyzed through the lens of Goffman's [1] work, and particularly, "The Presentation of the Self In Everyday Life". The profile is often regarded as a scene, while the action of sharing certain information becomes a way of performing.

Goffman's [1] metaphor of the dramaturgy of everyday life draws from the premise that individuals take up different roles in order to create an idealized version of their selves. These roles vary according to different contexts and according to what the audience expects as appropriate behaviour. In this context, he makes a distinction between "expressions given and expressions given off", where the latter consists of uncontrolled manifestations of the "true self". However, one key element in Goffman's[1] theory is how actions are bounded in space and time and oriented towards specific goals. Goffman [1] described these specific settings in terms of "front region" and the "back region". In the front stage, we are trying to present an idealized version of the self according to a specific role: to be an appropriate server, lecturer, audience member, and so forth. The back- stage, as Goffman [1] says, is "a place, relative to a given performance, where the impression fostered by the performance is knowingly contradicted as a matter of course" (p. 112).

The audience circumscribes those who observe a given actor and monitor his performance. More succinctly, these are those for whom one "puts on a front." This front consists of the selective details that one presents in order to foster the desired impression alongside the unintentional details that are given off as part of the performance. Moreover, a front involves the continual adjustment of self-presentation based on the presence of others. The key point here is that individuals put on specific fronts and modify said fronts because of the sustained observation of an audience.

2 GOFFMAN AND SOCIAL NETWORKING SITES

Goffman has often been used as a theoretical framework for the study of SNS's. By SNS's I mean sites defined by combination of features that allow individuals to (1) construct a public or semi- public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system[2].

A common idea of articles about SNS's is that individuals use this tool to employ impression management (or the selective disclosure of personal details designed to present an idealized self). Authors who use Goffman in this manner include: Boyd & Heer [3]; Lampe et al. [4], Hewitt and Forte[5], Lewis et al [6], Tufekci [7]. When regarding Facebook in this perspective, I find Hogan's[8] critique to be of utmost significance. He discusses the dichotomy between performance as an ephemeral act and recorded performance and points out that a recorded performance can be taken out of its original context and be played in another setting. He argues that everyday life is now replete with reproductions of the self and those reproductions lack the aura of the original, just as it happens with artwork. Thus, he introduces the exhibitional approach, which is specific to sites where users are not necessarily copresent in time. These sites require a third party to store data for later interaction, which places the analysis in a different zone from the focus of Goffman's work.

A second important distinction Hogan[8] makes is between information which is addressed and information which is submitted. On SNS's, the information shared is not bound to a specific audience. Computers take up the function curators have in an art exhibition, while users are equated to artefacts, as they can be filtered and searched.

[1] Information and Knowledge Society doctoral candidate at the Internet Interdisciplinary Institute, Open University of Catalunya,, Roc Boronat 11, C.P. 08018 Barcelona, Email: igreti_iulia@uoc.edu

[2]

I consider Hogan's[8] analysis of the presentation vs. exhibition of the self in SNS's to be very accurate, as I also completely adhere to the distinction he points out between actor and artefact. Furthermore, I believe this distinction to have very deep implications on the construction of the self in and through online environments. If we conceptualize dramaturgical performance as a means for the presentation of the self, to what end does replacing the performance with an exhibition lead? I believe the crucial point of this shift, which Hogan[8] indirectly touches upon, is the objectification of the self. And, although fostered by the exhibition-like setting, this process of objectification has gone beyond the possibility to filter information or to search for certain individuals according to a series of criteria. It is now a mechanism that users have internalized and they put a certain effort into directing it towards a desired finished good. They are aware of the exhibition they are letting themselves be placed in, and are trying to determine their exact position through the information they share, or in other words, through the artefact they become.

3 IMPLICATIONS OF EXHIBITED SELVES

I believe that one of the main ways in which users expose themselves to the objectification that Facebook employs and at the same time contribute to it is through consumption. Extending F. de Saussure's linguistic structuralism, Baudrillard [9] argues that consumption is a way to differentiate ourselves socially, a result of the need not for a particular object due to its intrinsic value (as in classic Marxist theory), but a need for social difference and meaning. Some of the main components of users' profiles are related to their taste in music, movies or books. From this point of view, Facebook can also be seen as an accurate application of Bourdieu's theory of social distance. Each user interprets what he likes or what others like as indicators of, broadly speaking, social prestige. Bourdieu [10] explains: "Because different conditions of existence produce different habitus- systems of generative schemes applicable, by simple transfer, to the most varied arias of practice, the practices engaged by different habitus appear as systematic configurations of properties expressing the differences objectively inscribed in conditions of existence in the form of systems of differential deviations which, when perceived by agents endowed with schemes of perception and appreciation necessary in order to identify, interpret, evaluate their pertinent features, function as life styles." It is due to these systematic configurations that Facebook, and all SNS's for that matter, are so successful. Information about what one does around the clock, what books he reads, what movies he watches, who he talks to and what they talk about is not interesting in itself, but it becomes interesting as a tool for systematization. Moreover, I am skeptical of the explanation that comes at hand, about people finding pleasure in gossip. Even Dunbar's [11] grooming explanation of gossip as the human version of social grooming in primates seems to have limited applicability in the type of interaction social network sites host. Thelwall and Wilkinson [12] emphasize the difference between social grooming and information gathering, underscoring that social grooming requires maintaining relationships with others through gossip or other minor activities. They point out the fact that empirical evidence support more the hypothesis of pure information gathering rather than social grooming, as users commonly visit profiles unobtrusively, without communicating with the individuals they are gathering information on. Although a case can be made that creating a profile, regular posting or following other users activity is a form of forging bonds, affirming relationships, displaying bonds, and asserting and learning about hierarchies and alliances, the reduced dimensionality of "the other", the decontextualization and the accessibility of others in the absence of any form of interactivity bring an essential change to the initial premises. Consequently, I believe all of the cues shared in a profile are interpreted according to the user's own system of codes in a way that helps him create a unified artefact of the other. If selves are indeed, as argued by post-modernists, not coherent narratives, but disarticulated fragments that are often contradictory, than it's not difficult to understand why a simulated objectification of others that makes sense and can be placed on a social mapping sounds tempting for most of us.

However, individuals are not only exposed to this process, they are also aware of it and consciously engaging in it themselves. Consequently, an expected outcome is to artificially create a habitus that one predicts will result in them gaining a certain position in the social maps others create, which, ultimately, lies at the core of the objectified simulacrum of the self. In practical terms, symbolic fictions are replaced by simulations of capital through the hierarchization of the codes, or, in other words, the rating of preferences. Undoubtedly, the rating is strongly influenced by one's subjectivity, but even more so, by their constructed simulation of subjectivity. Parallel to their evaluations, Facebook users emphasize different dimensions on their own profile, they simulate a certain type of capital, but they always activate (or at least aim to activate) on a market with those with similar evaluations of certain codes. Furthermore, within the "market" created around each type of activity, there is a hierarchy of the products that can be consumed in order to maximize that experience. Just like there is a market of detergents where one consumes a product or the other according to the evaluation of their capacity to wash clothes, there is a market of adventurous trips where the most appreciated would be the trips to inaccessible, wild or dangerous places. But, on markets such as music or other arts, a hierarchy of products is very difficult to be obtained, due to the subjectivity implied in the evaluation of what maximizes the experience. And as subjectivity is strongly shaped by offline social class belonging, so is the system of codes according to which one establishes a hierarchy of music genres. What happens is that individuals with similar social status will have similar codes and will end up having similar preferences on markets that are not intrinsically related. Thus, what Facebook does is list most of the cues needed for an individual to be "read" as a whole according to a series of codes. That is one of reasons why I believe Facebook moves from the commodified structure of aspects of our lives to a transparent unified commodification of selves. The author of a profile doesn't just present his preferences or hobbies; he presents those preferences and hobbies that allow him to wrap himself up to the image he wants to obtain, assuming the viewer shares the same system of codes. And he is viewed as a holistic entity. Each new item a user posts is filtered through questions about how that information will contribute to the final object users want to make of themselves. Shifting from presentation to exhibition gives the user complete control over what one lets others see. But that doesn't mean it also gives control over what they perceive or interpret from your

exhibited self image. In the absence of non-verbal signs, what you give is still not the same as what you give off. When an individual is presenting himself to an audience, he plays the role he believes is expected to play in that particular context. But one of the consequences of the collapsing contexts that are often invoked when talking about online environments is that the expectancies are directed towards you as a whole. One judges a math teacher by his math knowledge, by his conduct during class, by his interactions with parents or peers, etc. and he might be able to present himself in a positive light. Yet, if the same teacher activates on an SNS where he makes spelling errors in his posts, the entire presentation is undermined. So, irrespective of how reliable the image created through an exhibition is in comparison to a role delivered in face to face contextual presentation, it is still going to be relevant in the evaluation of the audience, unless they already have a holistic judgement of the person in question.

Thus, the simulacrum of the self is simultaneously a resource for generating meaning and mapping the social space and the main outcome of the same process. However, I expect the limit of simulation to be generally reached at the point where it becomes impossible for the subject to compatibilize it with his own self image. Thus, I am probably not willing to post photo shopped pictures of me in the Amazonian Jungle, although I haven't left my home town in months, but I am willing to post it if I lived for a week in a nearing locality and I have just taken a 2 hour excursion to the wilderness. My explanation for that is the need not just for others, but also for the user to interpret his own signs in a way that would lead him to consider he is close to his socially constructed ideal self.

Going even further, Facebook is essentially consumerist due to its de-humanizing character. What happens is that friends on Facebook are not people one feels emotionally attached to, but opportunities to watch impersonal narratives. Just like the object of consumption, the user does not function via the utilitarian or the personal: it functions via its relations with other objects.

Another essential difference between the presented self and the exhibited artefact representing the self is the final purpose of the presentation. Although both are often judged in terms of "impression management" there are important nuances that need to be distinguished. When presenting one's self in the front stage, an individual is strongly conditioned by issues of adequacy between his actions and the role he/ she is assuming rather than by identity matters. However, when creating an exhibitionary space of your self, the user is anticipating and aiming for a global evaluation. Questions of what a teacher or a waiter are expected to do are replaced by questions of who one is or what he/she is like. Above, I have talked about the use of Facebook for social mapping. When creating a traditional presentation of the self, it often happens that individuals don't expect to be mapped according to it and they often are not. Compatibility of one's behaviour with what he/she believes the audience expects is less revealing in terms of symbolic capital than creating a profile on SNS's would be. One key indicator for symbolic capital that is absent in everyday interactions is the selection of relevant information that is supposed to reach an audience. In face to face interactions, the selection is, at least partly, given by the context, or by the role assumed, but in the virtual exhibition, the user accounts entirely for the decision on whether certain information is worth sharing.

But ultimately, face to face interacting, seen from the dramaturgical perspective, is most of the time a spectacle of masks, and the mask tells little about the actor. If someone is trying to evaluate the actor behind the mask, they will probably try to look beyond it, search for giveaways the actor lets slip. In SNS's, the premise is that the profile, the artefact is revealing of the self. When trying to learn more about some other user, someone does not look beyond the mask of the profile, but looks at it and through it. And sharing information you know is going to be viewed as representative of you determines the need for control. Zhu [13], for instance, says: 'people, despite their various cultural backgrounds, are believed to possess self-image/value and want their self-image/value to be appreciated and respected by other members of the community'.

On the other hand, this phenomenon has implications at the macro societal level. Qi [14] defines face as the social anchoring of self in the gaze of others and argues that the use of this concept in Chinese sociology can be related to Goffman's work. However, after discussing aspects about the universality of face and the relationship between a person's self-image and their social standing, the author shows concern over the "possibility of the reification of face, the generation of face as a conscious project of social relations(…)It is possible, then, that face considerations may go beyond a mere mechanism associated with social approval and disapproval of the thing that gives rise to face or subtracts from it, and that face itself becomes an object of self-conscious consideration. It is possible, then, that persons may be engaged in the construction of face as a self-conscious project, not only to achieve the pleasure of social approval and avoid the pain of social disapproval or censure, but also to engage in a politics of face as an explicit social practice." I believe this is no longer a danger, but a fact. Facebook is in itself a system that allows users to present information that they would share in every day social contact, while at the same time subtracting that information from any other purposeful interaction. Thus, the collective reification of selves results in a simulated sociality that is reduced to its political component.

Empirical evidence supporting this theoretical assumption can be found in Ledbetter et al. [15]. The article distinguishes between two essentially different uses of Facebook: online social connection and online self disclosure. In the context of this argument, I find that it is useful to focus on issues related to online self disclosure (OSD). One result compatible with my hypotheses is that OSD inversely predicted Facebook communication. Users who practice online social disclosure can be considered as having highly objectified selves, which translates into a stronger interest in the social mapping/ political positioning than in personal communication. Furthermore, online social connection emerged as a positive predictor for relational closeness, whereas online self disclosure was negatively associated with the same variable. Some might argue that the positive relation between online communication and relational closeness undermine the claim that Facebook is a means of dehumanizing selves. However, we need to keep in mind that the network of friends each user has is considerably larger than the number of close relations he/she has. So, we may expect the network to have a strengthening effect on existing strong ties, which, nevertheless, does not cancel the aspects of self objectification in relation to those with whom the user is not close. Moreover, we need to account for the fact that OSD and OSC usually coexist within the same account. Whether the user

communicates or not with some close friends over Facebook does not make him less exposed, or in Ledbetter et al.'s[15] terms, less self disclosed.

4 CONCLUSIONS

Facebook is one of the sites that foster the creation of personal profiles, where users submit data. Following the line of authors sustaining this activity contains an essential shift from traditional interpersonal communication, social grooming or the presentation of the self in Goffman's understanding, I explore the consequences this shift has on the construction of the self. Some of the main elements that distinguish SNS activity from other form of presence in the social life are absence of context, sharing information without direct communication, more control over what one shares (lack of non-verbal cues), the possibility to search for people, to filter them, to organize them according to certain criteria and so forth.

I argue that the voluntary enrolment in the practice of exposing in stead of presenting oneself results in the objectification of selves as artefacts and their consumption as narratives. From this point of view, the motivations behind willing reification I believe relate to gains in symbolic capital and upward mobility in the social field. And users find it straight forward to do so through the creation of a one-dimension self that allegedly meets the expectations of a reference group (or in some cases even individual) that the user strives to get closer to. Conversely, the monitoring of others appears to be a tool for elaborating the social map that surrounds the user, and is a necessary process for making an accurate estimation of the expectations of the reference group(s). When talking about exposure or reification of the self as a conscious action, the content that is exposed becomes a strategic, and thus extremely relevant, choice. Facebook users are aware of and seek to be evaluated by their posts, by what they share, by their likes and their guide to this construction is the habitus that their target group exposes. I expect this dynamics to lead to a simulacrum of self that is more than a front stage, because expectations are no longer related to specific roles, but to subjects as a whole and because of the underlying claim for authenticity.

Ultimately, what changes is the way in which we construct ourselves through and for others, as well as the mechanisms of evaluation others employ and those mechanisms interfere decisively with the core of all social relations. Therefore, I believe the analysis of reified selves is an important step within the broader thematic of the influence computer mediation has on subjectivities.

REFERENCES

[1] Goffman, E. *The presentation of the self in everyday life*, New York, NY: Anchor Books, (1959.)

[2] Boyd. d., & N. Ellison. *Social Network Sites: Definition, History, and Scholarship*. Journal of Computer-Mediated Communication, (2007) Retrieved from: http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html

[3]Boyd, D. & Heer, J. *Profiles as conversation: networked identity performance on friendster*, in Proceedings of the Hawai'i International Conference on System Sciences, Persistent Conversation Track, IEEE Computer Society, 4–7 January, Kauai, HI [4] J. Masthoff. Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers. *UMUAI,* 14:37-85 (2004).

[4] Lampe, C., Ellison, N. & Steinfield, C. *A familiar Face(book): profile elements as signals in an online social network*, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM Press, New York, pp. 435–444 (2007).

[5] Hewitt, A., & Forte, A. *Crossing boundaries: Identity man- agement and student/faculty relationships on the Facebook*. Poster session presented at CSCW, Banff, Alberta, Canada (2006).

[6] Lewis, K., Kaufman, J., & Christakis, N. *The taste for privacy: An analysis of college student privacy settings in an online social network*. Journal of Computer-Mediated Commu- nication, 14, 79-100 (2008).

[7] Tufekci, Z. *Grooming, Gossip, Facebook and Myspace*. *Information, Communication & Society*, *11*(4), 544-564 (2008).

[8] Hogan, B. The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online. *Bulletin of Science, Technology & Society*, *30*(6), 377-386 (2010).

[9] Baudrillard, J. *The Consumer Society: Myths and Structures*, Sage Publications (2003)

[10] Bourdieu, P. "Distinction- A Social Critique of the Judgment of Taste" (translated by Richard Nice), Harvard University Press (1983).

[11] Dunbar, R. *Grooming, Gossip, and the Evolution of Language*, Harvard University Press, Cambridge, MA (1998).

[12] Thelwall, M., & Wilkinson, D. Public Dialogs in Social Network Sites : What Is Their Purpose ? *Journal of the American Society for Information Science*, *61*(Cmc), 392-404 (2010).

[13] Zhu, H. 'Looking for Face', Journal of Asian Pacific Communication 13: 313–21(2003).

[14] Qi, X. Face: A Chinese concept in a global sociology. *Journal of Sociology*, *47*(3), 279-295 (2011).

[15] Ledbetter, a. M., Mazer, J. P., DeGroot, J. M., Meyer, K. R., & Swafford, B. Attitudes Toward Online Social Connection and Self-Disclosure as Predictors of Facebook Communication and Relational Closeness. *Communication Research*, *38*(1), 27-53 (2010).

Qualitative Methods of Link Prediction in Co-authorship Networks

Elisandra Aparecida Alves da Silva[1] **and Marco Túlio Carvalho de Andrade** [2]

Abstract. Link Prediction is useful in many application domains, including recommender systems, information retrieval, automatic Web hyperlink generation, and protein/protein interactions. In social networks it can be used for recommending users with common interests which is a useful mechanism to improve and to stimulate communication. This paper presents qualitative methods for link prediction in co-authorship networks, which are based on Fuzzy Compositions to predict new link weights between two authors adopting not only attributes nodes, but also the combination of attributes of other observed links. Using DBLP dataset we explore the used attributes and demonstrate that qualitative methods represent a satisfactory approach in this context.

1 INTRODUCTION

Nowadays, many databases are described as a linked collection of interrelated objects. The networks formed by such objects can be homogeneous, in which there is a single object type and link type or heterogeneous networks in which objects and links may be of multiple types. An example of heterogeneous network is the WWW (World Wide Web), and examples of homogeneous networks include co-authorship networks, which are used in this project.

The main aim of traditional data mining algorithms is to find patterns in a dataset characterized by a collection of independent instances of a single relation. However, the application of traditional statistical inference procedures which use independent instances can lead to inappropriate conclusions about data [14]. According to [17], a challenge in the Data Mining area is to deal with richly structured, heterogeneous data. In this way, the link features between objects need to be used to improve the accuracy of predictive models [10]. Some of these features are mentioned by [6]: the correlation between attributes of interconnected objects and the existence of links between objects which present similarities.

Link Mining refers to Data Mining techniques that explicitly consider the links in the development of predictive or descriptive models of interconnected data. The Link Mining tasks, according to the taxonomy shown by [10], are: object-related tasks (Ranking, Classification, Clustering and Object Identification), link-related tasks (Link Prediction) and graph-related tasks (Subgraph Detection, Graph Classification and Generative Models for Graphs). Link Mining is an emergent area that represents the intersection of different areas: Link Analysis, Web and Hypertext Mining, Relational Learning and Inductive Logic Programming, and Graph Mining.

The main aim of Link Prediction is to determine the existence of a link between two entities using object or link attributes. Link Prediction is useful in different application fields, such as recommendation systems, detection of links not observed in terrorist networks, protein interaction networks, prediction of collaboration between scientists and Web hyperlinks prediction [33].

This paper presents qualitative methods for link prediction considering context information in co-authorship networks. A systematic process to evaluate Link Prediction methods based on non-dichotomic metrics for data selection, determination of new links and evaluation of results is used.

This work is organized as follows: Section 2 presents important definitions, Section 3 presents an overview of the main methods of Link Prediction and Section 4 presents the used process. Finally, the application of the process and the results are presented. The next section deals with the important definitions for this work.

2 DEFINITIONS

In this paper we consider the use of co-authorship networks variables. Thus, it is important to present the definitions related to these networks.

2.1 Co-authorship Network

According to [34], a social network is formed by a set of actors and their relationships: family, friendships, work, etc.. These relationships may be associated with the context in which user interacts with others. [31] points out that social networks express the world in motion that, according to [19], is a not well understand world, since these networks connect people which interact with others to share information in a structure that is constantly evolving. The social structure favors the information sharing between network actors, but it is important that these relations be consolidated allowing the actors to know their partners to establish trusting relationships and ensure an efficient information sharing.

[1] Federal Institute of São Paulo, Department of Informatics Av. Francisco Samuel Lucchesi Filho, 770 12929-600 Bragança Paulista, SP - Brazil, email: elisandra@ifsp.edu.br

[2] University of São Paulo, Polytechnic School, Dept. of Computer Engineering and Digital Systems Av. Prof. Luciano Gualberto, travessa 3, 158 05508-900 São Paulo, SP - Brazil, email: marco.andrade@poli.usp.br

According to [9], "a social network is a graph where people or organizations are represented by nodes connected by edges, which can correspond to strong social relationships sharing some characteristic. Analysis of this graph structure, as statistical analysis of nodes and/or edges attributes may reveal important individuals/organizations relationships and special groups."

[17] presents an analysis of Link Prediction methods in Social Networks, and believes that as part of recent research in large complex networks and their properties, considerable attention has been given to the computational analysis of social networks structures in which nodes represent people and other entities in a social context, and links represent interactions, collaboration and influence between entities.

Here, the co-authorship networks are social networks in which nodes represent authors and links represent their coauthored publications. However, one can explore different relationships in these networks, such as author-conference [2] and author-word [26].

Having the definition of social networks and the relevant aspects related to co-authorship networks, the definition of interaction is shown.

Some authors do not distinguish interaction and interactivity, but there are those that relate interaction to human relationships and interactivity to the human-machine interface. In this work, interaction refers to links between authors of a coauthorship network or author-author interaction. Therefore, it differs from the definitions presented, which are focused on user-computer interaction.

Having the interaction definition adopted, it is necessary to define the users relevant characteristics. Several characteristics can be used for user representation, as his knowledge about the system, goals, history, experience, and their preferences [22].

In a co-authorship network, communication enables experience and knowledge exchange in a bi-directional manner, which is an interesting feature to allow a more active interaction between actors. Thus, information about publications and coauthors, which refer to users interaction can be adopted for user representation. Next subsection presents the context definition.

2.2 Actor Context

[7] define context as any information that characterizes the entity situation, which may be one person, a computing device, or an relevant object for user-application interaction. For context information specification and modeling, five dimensions were suggested by [1]:

- Who: Identification of individuals engaged in a specific task;
- Where: User location;
- When: Temporal information such as time spent on particular task;
- What: Task performed by user;
- Why: Intention, which allows to understand the motivation for some action.

[27] presents the following classification for context:

- Computational: network connectivity, communication cost, resources, etc..;
- User: their characteristics, user profile, location, people nearby, social situation, etc.;
- Physical: light, noise, temperature, etc..;
- Time: day, week, season, etc..

In this work, nodes represent authors and links the coauthored publications. Therefore, the focus is the link prediction between authors, which is related to the User Context, determined by the relationships established with other authors. Next section present some Link Prediction methods based on structural properties.

3 LINK PREDICTION

The main goal of Link Prediction is to predict the existence of a link between two entities using features of objects and other observed links. The basic approach of Link Prediction methods is the classification of all node pairs based on the graph proximity measure. The link weight called $score(x, y)$ is assigned to each node pair x and y, and then a list is generated in decreasing order of score. Considering node x, $\Gamma(x)$ denotes a neighbor set of x. Neighbors of x are the nodes which are directly connected with x.

Thus, these methods can be seen as the computation of proximity measure or similarity between nodes x and y, related to the topology of the network. In general, these methods derive from Graph Theory and Social Network Analysis and are designed to measure similarity between nodes. According to [17], these methods need to be modified for applications to different contexts.

Many approaches are based on the idea that the greater the number of common neighbors between two objects, the greater the chance of a link between x and y. [6] and [15] have proposed abstract models for network growth using this idea. These authors present the most direct idea of the application of Common Neighbors to Link Prediction. [21] used this measure in the context of collaborations network.

$$score(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

[3] used the proximity idea to verify the similarity between personal web pages. They assume common neighbors with lower degrees as more relevant, as follows:

$$score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log(|\Gamma(z)|)}$$

Another approach, called Preferential Attachment, assumes that the probability of a new link involving x and y is proportional to the number of their neighbor's links. This measure is given as follows [4]:

$$score(x, y) = |\Gamma(x)| \times |\Gamma(y)|$$

The Link Prediction methods shown above are based on structural properties of networks and do not consider the connection weights between users. [20] proposed some adjustments based on proximity measures to be used in online social networks. As user's personal information is not generally available in these networks, only the structural properties have been used. Additionally, the connection weight between x and y, $w(x, y)$, was defined as the number of encounters between x and y. A simple adaptation of Common Neighbors including link weights is presented in [20]:

$$score(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(y,z)}{2}$$

In this context, there are also approaches based on path analysis [16],[13] and on graph structure of networks [12]. Different approaches consider the application of probabilistic models [33] and similarity measurements between two objects [24], [30]. The main problems related to these approaches are the high complexity of the probabilistic models and the utilization of node attributes in similarity measurements that sometimes are not available in networks. Additionally, the link information is not considered in these approaches.

All the methods shown in this section are used for new link determination. However, when they are adopted or proposed, the tasks necessary to evaluate the link predictions are not identified. In general, a simple strategy is adopted for selecting a subset of data (network used to generate new links), which are used for result evaluation, as an example, nodes which have at least a determined number of edges are selected [17]. And to compare the results, in general, the ROC curve and/or related metrics are used [12]. The next section deals with the used process for link prediction evaluation.

4 USED PROCESS

The used process involves the tasks shown in [28]: Data Selection, New Link Determination and Result Evaluation. In the data selection task, the use of fuzzy sensors is considered, the determination of new links is based on fuzzy compositions and the fuzzy ROC curve and AUC are used to evaluate the results. Hence, we use a process based on non-dichotomic metrics in order to evaluate the methods of Link Prediction, which allows the use of specialist's knowledge and adopting a perspective more similar to the human perception of the problem.

The following sections present the methods used in the tasks identified.

4.1 Data Selection

According to [18], the aim of a sensor is to generate a symbolic linguistic representation from numerical measurements, i.e., a numeric-linguistic conversion considering the subjectivity of the problem.

Thus, the sensor creates a symbolic qualitative description in two stages: (1) numeric measurement and (2) numeric-linguistic conversion. A numeric measure, generally obtained by an electronic processing, provides an objective quantitative description of objects. The measure language, usually obtained from the interrogation of users provides a qualitative description of subjective objects. The conversion should provide a very accurate description as that performed directly by a human. Therefore, to implement a symbolic sensor, the symbolism of the adopted language should be considered in order to artificially reproduce the human perception of the measure.

The fuzzy sensor used for data selection includes two input variables: $NumberOfPapers$ and $NumberOfCoauthors$ and an output variable that determines the choice of the node. Input variables are thus called because they represent authors of a co-authorship network. However, they can be obtained in different areas. The $NumberOfPapers$ is the number of encounters with others users and the $NumberOfCoauthors$ represents the number of neighbors.

Next section presents the link prediction methods considering non-dichotomic metrics.

4.2 New Link Determination

This work views the link weight between two users x and y as the "relation quality". This measure is obtained by the application of approaches that use features from co-authorship social networks, which can be directly used in other domains.

Different approaches based on fuzzy theory are revealed here. These approaches consider the use of fuzzy compositions to determine new links between two authors and employ the relation quality to determine the link weight.

The approaches consider that the quality of the relation between two authors is higher in the following situations:

- when two authors have a large number of papers, mainly in recent years;
- when the average of coauthors of the authors in the relation is low, but the common coauthors are not considered as they influence the relation in a positive way.

The next sections present the technique used for new link determination.

4.2.1 Fuzzy Compositions

Supposing that $R(X,Y)$ and $S(X,Y)$ are two fuzzy relations [3], the composition $C(X,Z)$ between $R(X,Y)$ and $S(Y,Z)$ is a fuzzy relation now between X and Z, using Y as a bridge (transitivity) [35]. It is given by:

$$C(X,Z) = R(X,Y) \circ S(Y,Z)$$

Therefore, using relation compositions, it is possible to predict new link weights connecting users that are not yet connected. The operator used in this work is the Max-product.

4.2.2 Relation Quality

This measure represents the quality of the relation between two users. We adopt different approaches to obtain this value.

The input variables used are NumberOfPapers, CoauthorsAverage and RelationTime.

NumberOfPapers is the number of papers coauthored by A and B;

CoauthorsAverage is the average of coauthors of A and B, but the common coauthors are not considered. $\Gamma(A)$ is the number of coauthors of A and $\Gamma(B)$ is the number of coauthors of B. This value is obtained as follows:

$$Co = \frac{\Gamma(A) + \Gamma(B)}{2} - (\Gamma(A) \cap (B))$$

RelationTime is the difference between the last year of training and the year of the oldest paper.

[3] A fuzzy relation establishes associations of different truth degrees between related elements, which are similar to the Fuzzy Set membership degrees [35]. A fuzzy relation example is given by "physical similarity between members from x and y".

Some of used rules are revealed below:

if CoauthorsAverage is low AND NumberOfPapers is low AND RelationTime is low THEN RelationQuality is regular
if CoauthorsAverage is low AND NumberOfPapers is low AND RelationTime is high THEN RelationQuality is low

RelationTime is important in case of low coauthors average and low number of papers. In this situation, RelationTime is used to determine if quality is low or regular.

For experiments the combination of these variables was considered to analyze what is the better choice in the context of co-authorship networks.

The next section presents the method used to evaluate the results of Link Prediction methods.

4.3 Result Evaluation

According to [25], the ROC analysis is a graphical method to evaluate diagnostic and prediction systems. The ROC graphs were initially proposed to analyze the quality of signal transmission [8]. Nowadays, they are used as a powerful tool to evaluate classifiers in Machine Learning and Data Mining areas [5], [29]. The ROC curve is obtained from the rate of false positives and true positives. Hence, it is possible to compare these values in various cutoff points, not just considering a single threshold. A measure often used to evaluate classifiers is the Area Under the Curve, which can range from 0 to 1, and the greater the value, the better their performance. In Link Prediction context, the main goal is to determine the existence of a link between two entities. In order to do so, the link weight is assigned to each pair of nodes x and y, and then a list is generated in decreasing order of score. This value can represent the membership degree of the link to the fuzzy set *Positive*. Thus, given that the methods provide a value representing the weight of the link and not only its existence, one can use a fuzzy method to generate the ROC curve and evaluate the Link Prediction methods.

The Fuzzy ROC curve is used to evaluate the results of new link determination methods. The main advantage of this method is the adoption of non-dichotomic representations to the result of the new link determination method (predicted class) and/or the real class. To create the traditional ROC curve, a threshold is selected to the predicted class, making the values be binarized in *Positive* and *Negative*. The true class is determined by the presence or absence of the sample at the test base, also using a dichotomic representation.

The Fuzzy ROC curve used to evaluate the prediction of links adopts a non-dichotomic representation for the result of the new link determination method.

To create the Fuzzy ROC curve, it is necessary to define the fuzzy sets which represent the values to predicted and real class of a new link determined by a method. Thus, let X the instance set given by a new link determination method. The fuzzy subset P_t of X is the set of ordered pairs defined as:

$$P_t = \{(x, \mu_{P_t}(x)) \mid x \in X \; e \; \mu_{P_t} \rightarrow [0,1]\}$$

where $\mu_{P_t}(x)$ is the membership degree of x to the positive links P_t of true class.

And their complement is defined as:

$$\bar{P_t} = \{(x, \mu_{\bar{P_t}}(x)) \mid \mu_{\bar{P_t}}(x) = 1 - \mu_{P_t}(x)\}, \forall x \in X$$

where $\mu_{\bar{P_t}}(x)$ is the membership degree of x to the set $\bar{P_t}$ of negative links. To analyze the method performance, it is necessary to verify if the predicted class is positive or negative. Hence, the set *Positive* can be defined to the Predicted class as follow:

$$P_p = \{(x, \mu_{P_p}(x)) \mid x \in X \; e \; \mu_{P_p} \rightarrow [0,1]\}$$

where $\mu_{P_p}(x)$ denotes the membership degree of x to the set P_p. And so their complement is:

$$\bar{P_p} = \{(x, \mu_{\bar{P_p}}(x)) \mid \mu_{\bar{P_p}}(x) = 1 - \mu_{P_p}(x)\}, \forall x \in X$$

Knowing the membership degree for each instance of the subsets shown, the operators maximum and minimum defined by [23] can be applied to determine the membership degree of a case to each category. Thus, since: $TP = P_p \cap P_t$, $TN = \bar{P_p} \cap \bar{P_t}$, $FP = P_p \cap \bar{P_t}$,$FN = \bar{P_p} \cap P_t$.

The membership functions to each case are given as: $\mu_{TP}(x) = \mu_{(P_p \cap P_t)}(x)$, $\mu_{TN}(x) = \mu_{(\bar{P_p} \cap \bar{P_t})}(x)$, $\mu_{FP}(x) = \mu_{(P_p \cap \bar{P_t})}(x)$, $\mu_{FN}(x) = \mu_{(\bar{P_p} \cap P_t)}(x)$.

Thus,

$$\mu_{TP}(x) = min[\mu_{P_p}(x), \mu_{P_t}(x)], x \in X$$

$$\mu_{TN}(x) = min[\mu_{\bar{P_p}}(x), \mu_{\bar{P_t}}(x)] = min[1 - \mu_{P_p}(x), 1 - \mu_{P_t}(x)]$$

$$\mu_{FP}(x) = min[\mu_{P_p}(x), \mu_{\bar{P_t}}(x)] = min[\mu_{P_p}(x), 1 - \mu_{P_t}(x)]$$

$$\mu_{FN}(x) = min[\mu_{\bar{P_p}}(x), \mu_{P_t}(x)] = min[1 - \mu_{P_p}(x), \mu_{P_t}(x)]$$

$\forall x \in X$. Since $\mu_{TP} + \mu_{TN} + \mu_{FP} + \mu_{FN} = 1$. To generate the fuzzy ROC curve, the values of true positives and false positives rates can be obtained from the measurements *Sensitivity* and $Specificity$ also associated with the ROC graph:

$$Sensitivity(\mu_P(x)) = \frac{\sum \mu_{TP}(x_i)}{\sum \mu_{TP}(x_i) + \sum \mu_{FN}(x_i)}$$

$$Specificity(\mu_P(x)) = \frac{\sum \mu_{FP}(x_i)}{\sum \mu_{TN}(x_i) + \sum \mu_{FP}(x_i)}$$

$\forall i, i = 1, 2, ..., n$ where x_i is the i-th case of the sample set and n is the total number of cases. The ROC curve is generated using fuzzy rate of false positives ($Specificity$) and true positive (*Sensitivity*).

4.4 Differential Aspects

The innovative aspects are shown below:

- Use of non-dichotomic metrics for the new link determination;
- Fuzzy composition is used to predict new link weights, considering:
  - Utilization of both objects attributes and link features to determine the relation weight, which is called Relation Quality. The use of objects' attributes is accomplished by the adoption of the following measures: (1) Average of coauthors of the users present in the relation and (2) Common neighbors.
  - The utilization of link features is obtained by the use of the measures: (1) Number of coauthored papers, which represents the number of encounters of the users and (2) Relation time.

- Fuzzy AUC associated with Fuzzy ROC Curve is used to evaluate the results of Link Prediction.

5 EXPERIMENTS

Suppose we have a social net $G = \langle V, E \rangle$ in which edge $e = \langle u, v \rangle \in E$ represents the total of interactions (co-authored papers) between u and v at different times. Having times t_0, t'_0, t_1, t'_1, and assuming that $t_0 \text{ ¡ } t'_0 \text{ ¡ } t_1 \text{ ¡ } t'_1$. $[t_0, t'_0]$ the training interval and $[t_1, t'_1]$ the test interval are considered. Let $G[t, t']$ consist of all edges in t and t'. Thus, an algorithm has access to network $G[t_0, t'_0]$, and generates a list of links that are not present in $G[t_0, t'_0]$ which needs to be verified in $G[t_1, t'_1]$.

[17] observed that the evaluation of Link Prediction methods use parameters $k_{training}$ and k_{test} and assumed that the $Core$ set are nodes that belong to at least $k_{training}$ links in $G[t_0, t'_0]$ and at least k_{test} links in $G[t_1, t'_1]$. In our work, we consider set as nodes selected by the use of Fuzzy Sensors which consider the variables $NumberOfPapers$ and $NumberOfCoauthors$.

The training interval is denoted as $G_{collab} = \langle A, E_{old} \rangle$ and E_{new} is used to denote the link set $\langle u, v \rangle$, u and $v \in A$. Let u and v co-author a paper during the test interval but not in the training interval ($E_{new} = A \times A - E_{old}$). These are the new interactions sought to be predicted.

Each link predictor p produces a list L_p of pairs $A \times A - E_{old}$ in decreasing order of score. For the evaluation, we focus on the $Core$ set, thus we denote $E^*_{new} := E_{new} \cap (Core \times Core)$ and $n := |E^*_{new}|$. Thus, the first n pairs in the list L_p in $Core \times Core$ are considered to determine the Area Under Curve.

The experiments were performed according to the process presented in the previous section. The DBLP dataset is shown in the next section.

5.1 Dataset and Setup

DBLP (Digital Bibliography & Library Project) is the dataset used in the experiment. This dataset contains data of Computer Science publications and has been used in different works [11], [33].

The DBLP Computer Science Bibliography from the University of Trier contains more than 1.15 million records. DBLP contains details from publications of conference proceedings related to Data Mining, Databases, Machine Learning, and other areas. The dataset is public and is in XML format [32].

6 RESULTS

The fuzzy methods proposed in this paper must be evaluated by comparing their performance measures. Table 1 shows the information about dataset and Table 2 presents the additional information about dataset.

Table 1. Training and Test

Period	Train.	Test	$\|E_{old}\|$	$\|E_{new}\|$	$\|E^*_{new}\|$
1	1999-2004	2005-2007	695906	852388	131904
2	2001-2006	2007-2009	1027172	1040058	191390

Table 2. Number of Authors and Papers

Period	Aut. Train.	Pap. Train.	Aut. Test	Pap. Test
1	323118	404432	368542	384311
2	426631	546927	443320	450703

Table 3. Traditional AUCs

Period	C	P	T	C+T	P+T	C+P	C+P+T
1	0.503	0.581	0.544	0.585	0.577	0.599	0.605
2	0.486	0.578	0.559	0.582	0.578	0.607	0.606

Table 3 presents the AUCs in both period, where C is $CoauthorsAverage$, P is $NumberOfPapers$ and T is $RelationTime$. The AUCs indicate that the use of all variables presents the best performance in period 2. The use of one variable shows that $NumberOfPapers$ is better than others in both period and $CoauthorsAverage$ is the worst. The use of two variables indicates that $CoauthorsAverage$ combined with others variables is the best approach. The use of just one variable presents worse results in period 1 and 2. In this case, the variable $NumberOfPapers$ presents the best performance.

The traditional and Fuzzy AUCs shown very similar results, but when using $CoauthorsAverage$ and $RelationTime$ $(C + T)$, $CoauthorsAverage$ and $NumberOfPapers(C + P)$ or three variables$(C + P + T)$, Fuzzy AUC can detect some variations not considered by traditional AUC.

7 CONCLUSION

The results show that when using one variable the $NumberOfPapers$ is better than others in both period and $CoauthorsAverage$ is the worst. The use of two variables indicates that $CoauthorsAverage$ combined with others is better than others approaches, but the use of just one variable presents the worst results in both periods. In this case, the variable $NumberOfPapers$ revealed the best performance.

A process based on non-dichotomic metrics in order to evaluate the Link Prediction methods allows the use of the specialist's knowledge and adopting a perspective more similar to the human perception of the problem. The use of a fuzzy model to determine the $RelationQuality$ is interesting because it allows the specialist knowledge in the field to be exploited in the definition of variables, since some features are particular to that type of social network. In important works [3, 17] only one variable (common authors number) is considered, then the results indicate that there are variables related to the context that can be better explored in link prediction methods.

Table 4. Fuzzy AUCs

Period	C	P	T	C+T	P+T	C+P	C+P+T
1	0.503	0.587	0.545	0.594	0.582	0.616	0.619
2	0.486	0.579	0.563	0.594	0.578	0.623	0.621

REFERENCES

[1] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles, 'Towards a better understanding of context and context-awareness', in *Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing*, pp. 304–307, London, UK, (1999). Springer-Verlag.

[2] Evrim Acar, Daniel M. Dunlavy, and Tamara G. Kolda, 'Link prediction on evolving data using matrix and tensor factorizations', in *ICDMW '09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, pp. 262–269, Washington, DC, USA, (2009). IEEE Computer Society.

[3] Lada A. Adamic and Eytan Adar, 'Friends and neighbors on the web', *SOCIAL NETWORKS*, **25**, 211–230, (2001).

[4] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, 'Evolution of the social network of scientific collaborations', *Physica A: Statistical Mechanics and its Applications*, **311**(3-4), 590 – 614, (2002).

[5] A. P. Bradley, 'The use of the area under the roc curve in the evaluation of machine learning algorithms', *Pattern Recognition*, **30**(7), 1145–1159, (1997).

[6] Jorn Davidsen, Holger Ebel, and Stefan Bornholdt, 'Emergence of a small world from local interactions: Modeling acquaintance networks', *Physical Review Letters*, **88**(12), 128701, (March 2002).

[7] Anind K. Dey, Gregory D. Abowd, and Daniel Salber, 'A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications', *HCI Journal*, **16**, 97–166, (2001).

[8] J. P. Egan, *Signal detection theory and ROC analysis*, Academic Press, New York, USA, 1975.

[9] Carla M. D. S. Freitas, Luciana P. Nedel, Renata Galante, Luís C. Lamb, André S. Spritzer, Sérgio Fujii, José Palazzo M. de Oliveira, Ricardo M. Araújo, and Mirella M. Moro, 'Extração de conhecimento e análise visual de redes sociais', *Seminário Integrado de Software e Hardware. XXVIII Congresso da Sociedade Brasileira de Computação.*, (Julho 2008).

[10] Lise Getoor and Christopher Diehl, 'Link mining: A survey', *SigKDD Explorations Special Issue on Link Mining*, **7**(2), (december 2005).

[11] Mohammad Hasan, Vineet Chaoji, Saeed Salem, and Mohammed J. Zaki, 'Link prediction using supervised learning', (April 2006).

[12] Zan Huang, 'Link prediction based on graph topology: The predictive value of the generalized clustering coefficient', *Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (LinkKDD2006)*, (August 2006).

[13] Glen Jeh and Jennifer Widom, 'Simrank: a measure of structural-context similarity', in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pp. 538–543, New York, NY, USA, (2002). ACM.

[14] David Jensen, 'Statistical challenges to inductive inference in linked data', *Seventh International Workshop on Artificial Intelligence and Statistics*, (1999).

[15] Emily M. Jin, Michelle Girvan, and Mark E. J. Newman, 'The structure of growing social networks', *Physical Review E*, **64**(4), 046132, (2001).

[16] Leo Katz, 'A new status index derived from sociometric analysis', *Psychometrika*, **18**(1), 39–43, (March 1953).

[17] David Liben-Nowell and Jon Kleinberg, 'The link-prediction problem for social networks', *Journal of the American Society for Information Science and Technology*, **58**(7), 1019–1031, (May 2007).

[18] Gilles Mauris, Eric Benoit, and Laurent Foulloy, 'Fuzzy symbolic sensors-from concept to applications', *Measurement*, **12**(4), 357–384, (1994).

[19] José Luis Molina and Claudia Aguilar. Redes sociales y antropología: un estudio de caso (redes personales y discursos étnicos entre jóvenes en sarajevo). LARREA, 2005.

[20] Tsuyoshi Murata and Sakiko Moriyasu, 'Link prediction based on structural properties of online social networks', *New Generation Comput.*, **26**(3), 245–257, (2008).

[21] M. E. J. Newman, 'The structure of scientific collaboration networks.', *Proceedings of the National Academy of Sciences USA*, **98**(2), 404–409, (January 2001).

[22] L. A. M. Palazzo, 'Sistemas de hipermídia adaptativa', *XXI Jornada de Atualização em Informática. XXII Congresso da SBC. Florianópolis*, (Julho 2002).

[23] Raja Parasuraman, Anthony J. Masalonis, and Peter A. Hancock, 'Fuzzy signal detection theory: Basic postulates and formulas for analyzing human and machine performance', *Human Factors*, **42**(4), 636–659, (2000).

[24] Alexandrin Popescul and Lyle H. Ungar, 'Statistical relational learning for link prediction', *Workshop on Learning Statistical Models from Relational Data at IJCAI*, (2003).

[25] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, 'Curvas roc para avaliação de classificadores', *Revista IEEE América Latina*, **6**(2), 215–222, (JUNE 2008).

[26] P. Sarkar, S. M. Siddiqi, and G. J. Gordon, 'A latent space approach to dynamic embedding of co-occurrence data', *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AI-STATS)*, (2007).

[27] B. N. Schilit, *A Context-aware System Architecture for Mobile Distributed Computing*, Ph.D. dissertation, Columbia University, 1995.

[28] E. A. A. Silva and M. T. C. Andrade, 'A process based on the fuzzy set theory for evaluation of link prediction methods', in *Combined Proceedings of the Social Network Analysis and Norms for MAS Symposium - SN-MAS2010 Section at the AISB 2010 Convention*, pp. 16–21, De Montfort University, Leicester, UK, (2010). SSAISB.

[29] K. A. Spackman, 'Signal detection theory: Valuable tools for evaluating inductive learning', *Proceedings of the 6th Int Workshop on Machine Learning*, 160–163, (1989).

[30] Ben Taskar, Ming fai Wong, Pieter Abbeel, and Daphne Koller, 'Link prediction in relational data', *Neural Information Processing Systems*, (2004).

[31] Maria Inês Tomaél. Redes sociais: posições dos atores no fluxo da informação. R. Eletr. Biblioteconomia, 2006.

[32] University Trier. Digital bibliography & library project (dblp), 2009.

[33] Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy, 'Local probabilistic models for link prediction', in *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pp. 322–331, Washington, DC, USA, (2007). IEEE Computer Society.

[34] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, ENG and New York, 1994.

[35] L. A. Zadeh, 'Fuzzy sets', *Information and Control*, **8**(3), 338–353, (June 1965).

From Linguistic Innovation in Blogs to Language Learning in Adults: What do Interaction Networks Tell us?

Michał B. PARADOWSKI[1], Chih-Chun CHEN[2], Agnieszka CIERPICH[1], Łukasz JONAK[3]

Abstract. Social networks have been found to play an increasing role in human behaviour and even the attainment of individuals. We present the results of two projects applying SNA to language phenomena. One involves exploring the social propagation of neologisms in a social software (microblogging service), the other investigating the impact of social network structure and peer interaction dynamics on second-language learning outcomes in the setting of naturally occurring face-to-face interaction. From local, low-level interactions between agents verbally communicating with one another we aim to describe the processes underlying the emergence of more global systemic order and dynamics, using the latest methods of complexity science.

In the former study, we demonstrate 1) the emergence of a linguistic norm, 2) that the general lexical innovativeness of Internet users scales not like a power law, but a unimodal, 3) that the exposure thresholds necessary for a user to adopt new lexemes from his/her neighbours concentrate at low values, suggesting that—at least in low-stakes scenarios—people are more susceptible to social influence than may erstwhile have been expected, and 4) that, contrary to common expectations, the most popular tags are characterised by high adoption thresholds. In the latter, we find 1) that the best predictor of performance is reciprocal interactions between individuals in the language being acquired, 2) that outgoing interactions in the acquired language are a better predictor than incoming interactions, and 3) not surprisingly, a clear negative relationship between performance and the intensity of interactions with same-native-language speakers. We also compare models where social interactions are weighted by homophily with those that treat them as orthogonal to each other.

1 LANGUAGE PHENOMENA EXHIBITING COMPLEX SYSTEM CHARACTERISTICS

Within an individual, many linguistic mechanisms are at work, such as the perceptual dynamics and categorisation in speech, the emergence of phonological templates, or word and sentence processing. There are also a multitude of interactions simultaneously occurring at the society level between systems that are inherently complex in their own right, such as variations and typology, the rise of new grammatical constructions, semantic bleaching, language evolution in general, and the spread and competition of both individual expressions, and entire languages. Nearly two hundred papers have already been published dealing with language simulations. However, many of them, devoted to phenomena such as language evolution, language competition, language spread, and semiotic dynamics, were based on regular-lattice *in silico* experiments and as such are grossly inadequate, especially in the context of the 21st c. The models:

- only allow for Euclidean relationships (while nowadays more and more of our linguistic input covers immense distances; spatial proximity ≠ social proximity),
- are 'static' (while mobility is not exclusively a 20th or 21st-c. phenomenon, as evidenced by warriors, refugees, missionaries, or tradespeople),
- assume an identical number of 'neighbours' for every agent (4≤8),
- presuppose identical perception of a given individual's prestige by each of its neighbours[4], as well as
- invariant intensity of interactions between different agents,
- most fail to take into account multilingual agents[5],
- have no memory effect, and
- zero noise (while noise may be a mechanism for pattern change).

To address these limitations, rather than take a modelling outlook, we can start with analysing language phenomena in social networks—either by tapping into already available repositories of data nearly perfectly suited to large-scale dynamic linguistic analyses, such as the Internet, or by analysing communities of speakers via offline approaches—and subsequently applying SNA and other complexity science tools to the analyses. Roman Jakobson remarked already half a century ago on the "striking coincidences and convergences between the latest stages of linguistic analysis and the approach to language in the mathematical theory of communication" ([17] p. 570).[6]

[1] Inst. of Applied Linguistics, Univ. of Warsaw, ul. Browarna 8/10, 00-311 Warsaw, Poland. Email: michal.paradowski@uw.edu.pl; a.cierpich@student.uw.edu.pl.

[2] Centre d'analyse et de mathématique sociales – UMR 8557, École des hautes études en sciences sociales, 190-198, avenue de France, 75244 Paris cedex 13, France. Email: c.chen@abmcet.net.

[3] National Library of Poland, Al. Niepodległości 213, 02-086 Warsaw, Poland. Email: lukasz@jonak.info.

[4] But see e.g. [13] or [33] incorporating complex network architectures and differences in prestige.

[5] But see e.g. [2].

[6] « II est un fait que les coïncidences, les convergences, sont frappantes, entre les étapes les plus récentes de l'analyse linguistique

2 LANGUAGE ON THE INTERNET

Erstwhile research on language evolution and change focused on large time-scales, typically spanning at least several decades. Nowadays, observable changes are taking place much faster. According to [12] a new English word is born roughly every 98 minutes (admittedly an overrated estimate owing to methodological problems). Particularly useful for multi-angle analyses of language phenomena are Web 2.0 services, with content (co)generated by the users, especially the ones which allow enriching analyses with information concerning the structure of the connections and interactions between the participating users. This unprecedented reliance on news delivered by the users is also increasingly being observed in editorial offices and television newsrooms.

The uptake of novel linguistic creations in the Internet has been commonly believed to reflect the focus of attention in contemporary public discourse (suffice it to recollect the dynamics and main themes of status updates on Twitter following the presidential elections in Iran, Michael Jackson's death, Vancouver Olympic Games, and the recent Oscar gala, last July's L.A. earthquake, the Jasmine Revolution—by some also called the "Internet Revolution"—in Tunisia, the developments in Libya, the 2011 Tōhoku earthquake and tsunami, or ibn Laden's death, see e.g. [11]). However, even where the topics coincide, the proportions in the respective channels of information are divergently different (correlation at a level of a mere .3; e.g. [27], just as television ratings cannot be used to predict online mentions; [26]), just as not infrequently the top stories in the mainstream press are markedly different than those leading on social media platforms (e.g. [29]). The emotive content of comments on different social platforms is also distinctly different ([5], [6]).

Table 1. The microblogging site in numbers (at time of data dump)

Users	20k, over half logging on daily
Users in the giant component	5.5k (density 0.003)
Relations	110k
Tags[7]	38k
Tagged statuses	720k

While there does exist some scarce research looking at the emergence and spread of online innovation[8], studies that do so utilising social network data are next to non-existent. Our empirical research project has set out to investigate how mutual communication between Internet users impact the social diffusion of neological tags (semantic shortcuts) in Polish microblogging site Blip (for site statistics, see Table 1).

et le mode d'approche du langage qui caractérise la théorie mathématique de la communication. » (*Essais de linguistique générale*, 1967:87)

[7] By tags (or 'hashtags') we mean expressions prefixed with the number sign '#' and usually used in microblogging sites to mark the message as relevant to a particular topic of interest, or 'channel'.

[8] *Cf.* e.g. [24] for how the use of Internet chatrooms by teenagers is resulting in linguistic innovation within that channel of virtual communication, [18] for a discourse-analytic glance at the social practices of propagating online memes, or [22] for a visualisation of the 'competition' between top quotes in the news during the 2008 US presidential election.

3 TAGS AND SOCIAL COORDINATION

The intended purpose of tagging systems introduced to various Web 2.0 services was to provide ways of building *ad hoc*, bottom-up, user-generated thematic classifications (or "folksonomies"; [35]) of the content produced or published within those systems.

However, the tagging system of Blip became much more than that, as users redefined the meaning and modes of using tags. In the site, tagging is not merely a mechanism for retrospective content classification, but also provides institutional scaffold for on-going communication within the system. From the point of view of *individuals*, using a tag within a status update still provides information about what the update is about, but also implies joining the conversation defined by the tag, and, consequently, subscribing to the rules and conventions governing conversation. In this sense, the system of tags can be thought of as an institution (as sociologically understood), regulating and coordinating social conduct – here, mostly communication. From the *systemic* point of view, tags-institutions define what Blip.pl is about, the meaning of its dynamics, and its culture.

4 THE LONG TAIL OF THE BLIP CULTURE

One of the preliminary results obtained from the data analysis carried out concerns tag popularity, whose distribution scales like a power law (Fig. 1), a feature Blip shares with a wide range of natural, technological and socio-cultural phenomena (*cf.* e.g. [3], [25]). Our assumption is that at least a considerable proportion of popular Blip tags constitute the "meaning" and structure of the system, its cultural and institutional establishment, while the long tail consists of more or less contingent representations. Our interests lie in answering questions about the mechanisms which were responsible for the system becoming the way it is in terms of cultural tag composition.

Figure 1. Tag popularity distribution in Blip

5 SOCIAL INFLUENCE AND DIFFUSION

The most important mechanism we are looking for has to do with diffusion of innovation. Diffusion and creation of novelty has been traditionally assumed to be among the most important social processes [7]. In our case, each of Blip's tags,

a potential communication coordinator, had been first created by a user, then spread throughout the system with greater or smaller success (see Fig. 2). Some of the most successful, most frequently imitated tags have become Blip's culture and structure.

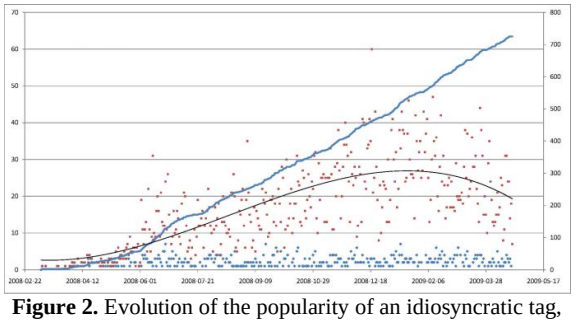


Figure 2. Evolution of the popularity of an idiosyncratic tag, relative to system size; abscissæ: time, ordinates left: percentage of saturation; ordinates right: absolute count; blue rhomb dots: first usages; red square dots: subsequent usages; thin black line: subsequent usage trend (multinomial); thick blue line: first usages cumulative

There are a number of theories explaining the mechanisms of diffusion of novelty, and one of our goals is to find out which best accounts for our data. Memetic theory assumes that ideas (here coded as words-tags) are like viruses which "use" the mechanisms of the human mind to reproduce. The most successful reproducers would be those optimally adapted to the environment of the mind – its natural dispositions and the ecosystem of already established ideas ([4], [8]).

The theory of social influence constructs a situation in which individual behaviour (including adoption of innovation) is contingent on peer pressure. The threshold model of collective behaviour postulates that a person will adopt a given behaviour only after a certain proportion of the people s/he observes have already done the same. This proportion—the "adoption threshold"—constitutes the individual characteristic of each member of the group ([14], [34]).

A third point of view is offered by the social learning theory [1], which assumes that innovation or behaviour adoption is a result of a psycho-cognitive process which involves evaluation of other people's behaviour and its consequences. In this case the adoption process is perceived as more reflexive and less automatic than the previous two ([15], [30]).

The preliminary analysis conducted involved calculating thresholds for all tag adoptions (i.e., their *first* usages). We describe the user-tag network with a bipartite graph $G = G(U,X,E)$, where U is the set of users, X is the set of tags, and E represents the edges between users and tags. The user-user network we define using a directed graph $D = D(U,H)$, where H is the set of edges. To every $e_{u\to x} \in E$ edge connecting user u to tag x added in time $\tau_{u\to x}$ we assign a variable $a(e_{u\to x})$, such that

$$a(e_{u\to x}) \begin{cases} 1 \text{ if in time } \tau_{u\to x} \text{ there is a neighbour of } u \text{ who is already connected to tag } x, \\ 0 \text{ else} \end{cases}$$

We capture the adaptive behaviour of a user with the statistical variable $\alpha_u \in \langle 0,1 \rangle$

$$\alpha_u = \frac{\sum_{e_{u\to x}\in E(u)} a(e_{u\to x})}{|E(u)|}$$

where $E(u) \in E$ is the set of connections of user u. A low value of α_u means that the user tends to introduce more innovation into the system.[9]

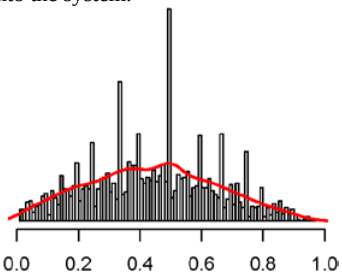


Figure 3. Creativity distribution in the microblogging site

Using the above notation, β_u is the (mean) measure of the number of alters (neighbours == followed users in Twitter/Blip terms) who had adopted a given tag before user u. We only consider first usages:

$$\beta_u = \frac{\sum_{e_{u\to x}\in E(u)} \frac{A_{(e_{u\to x})}}{H_{(t)}(u)}}{|E(u)|}$$

where:

- $A(e_{u\to xt})$ is the number of neighbours of u who are already connected to x at time $\tau_{u\to x}$ (in other words, it says how 'mainstream' the tag is);
- $H_{(t)}(u)$ is the number of neighbours of u at time t;
- $E(u)$ is the total number of (unique) tags used by u.

Thus, a high value of β_u corresponds to the user being more likely to be influenced by his/her neighbours.[10]

The resultant distribution of the thresholds is considerably skewed, with a median of 0.11 and a long tail of higher values (Fig. 4)[11]. This suggests that the population of Blip users is generally innovative and/or corroborates the viral model of diffusion over the two alternative theories mentioned above. However, we expect other factors (such as tag and user characteristics) to play an important role as well, especially since, contrary to many common expectations, expressions' popularity correlates negatively with low thresholds (Fig. 5).

An alternative explanation may be the classical diffusion process with population division into early adopters and laggards: thresholds rise with tags' popularity because users with lower thresholds had adopted them earlier (when the expressions were not yet popular). Our aim is to consider models that include these factors in explaining diffusion

[9] Although a large alpha can also be observed in cases where a user is surrounded by many neighbours who adopted a tag before her/him. Naturally, given the nature of the data recorded by social software, it is impossible to determine which entries a given user has actually read. This of course means that the posts published by 'followed' persons are merely treated as a realistic proxy of the data actually seen by the user.

[10] A thematic breakdown of the tags might reveal that humans succumb to influence more easily in certain contexts than others.

[11] The "humped" feature of the distribution tail stems from the skewed distribution of the variables used to calculate the threshold values.

mechanisms.

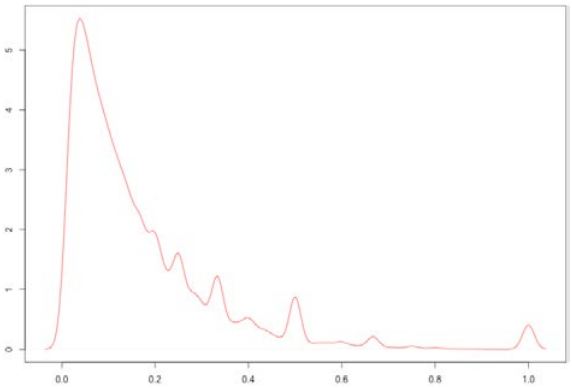


Figure 4. Distribution of tag adoption thresholds in Blip

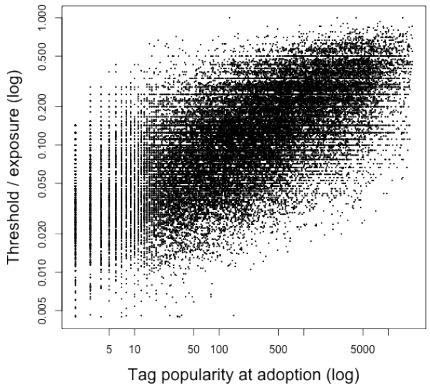


Figure 5. Relationship between tag popularity and exposure threshold

6 FOREIGN LANGUAGE STUDIES AND SOCIAL INTERACTION

In the field of foreign language studies, the past two decades have witnessed a significant increase in theories and research focused on the role of social interaction (e.g. socio-cultural theory [20], language socialisation hypothesis [19], or conversation analysis [9], [10]). These developments conceive of language learning as a process anchored in and configured through the activities in which the language user engages as a social agent [28]. Yet, to date no data-driven analysis has been carried out to investigate the impact of social network structure and peer interaction dynamics on second-language learning outcomes in the setting of naturally occurring face-to-face interaction.

7 SECOND LANGUAGE ACQUISITION AND LANGUAGE LEARNER NETWORKS: PARTICIPANTS, METHODS & MEASURES

During the 2010/11 academic year, a striking observation was made independently by several German-language instructors at one university in Baden-Wurttemberg: for the first time in a long while the cohort of Erasmus exchange students arriving at the university became a visibly cohesive group. This had a measurable impact on the improvement of their linguistic competence over the course of the academic year.

All members of the group (n=39) were approached with in-depth structured interviews, with the objective to grasp: (i) the precise individual, social and interactional factors impacting the acquisition process; (ii) the way in which language development is affected by the dynamics of peer interaction, and (iii) the impact of social network topology on motivation and learning outcomes. From these interviews, we were able to gain insight into the motivations, preferences and peer interaction among the participants. The goal was then to determine how, if at all, these were associated with performance. Because the number of participants was very low and the majority improved by one level, we chose to focus on over- and underperformers (improvement by two levels or no improvement) to try to identify the features and conditions that might explain their outcomes.

We measured *performance* in terms of self-reported improvement, taking the difference between the participant's initial level in German and their level at the end of the course.

Interaction frequency was assessed by the participants themselves and rated on a scale between 1 and 10, where a score of 10 was given for participants with which the individual felt s/he interacted most frequently.

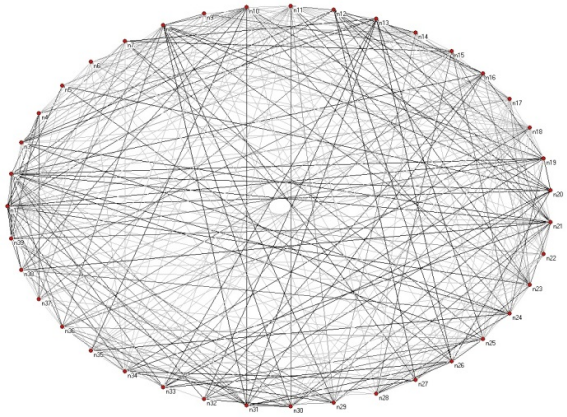


Figure 6. Bidirectional interactions in German; edge intensity indicates relative link weight

In our analyses, we consider six different weighted interaction networks, namely those of: (i) incoming interactions, where an individual *i* has an in-link from individual *j* if *j* has reported interacting with *i* (irrespective of whether or not *i* has reported such interaction); (ii) outgoing interactions, where individual *i* has an out-link to an individual *j* if *i* has reported interacting with *j*; (iii) the sum of general interactions; (iv) bidirectional interactions only; (v) incoming interactions *in German*; (vi) outgoing interactions *in German*; (vii) the sum of *German* interactions; (viii) bidirectional interactions *in German* (a snapshot of the last network is visible in Fig. 6).

The interactions were all normalised with respect to participants' general interactions (so, for example, if a participant had a high level of interaction, a score of 4 will be

treated the same as a score of 2 for a participant who did not interact very much).

Due to the low number of participants and the fact that the majority improved by one level, we had to ensure that any apparent similarities between strongly linked individuals (large frequencies of interactions) were not simply due to homogeneity. To address this, we compared the predictions that would be made by the network with those that would be made by the network randomly rewired. Rather than use traditional network analysis methods that depend on large numbers of nodes and links, we tested hypotheses by evaluating alternative models that overlay or weight networks. For example, to gain further insight on the interplay between social factors, language factors, and homophily ([21], [23]), we compare models where social interactions are weighted by homophily with those that treat them as orthogonal to each other.

8 SOCIAL INTERACTION AND PERFORMANCE

Using this multi-layered-network perspective to study socially distributed learning, we found:

(i) No direct association between outgoing interactions (neither general nor in German) and performance. However, when the outgoing *German* interactions were framed in the context of the *general* outward interactions (i.e., using $\frac{s_{german}}{s_{general}}$, indicating the degree to which they interacted in German less or more when compared with their general interactions), there appeared to be a positive association (see Fig. 7);

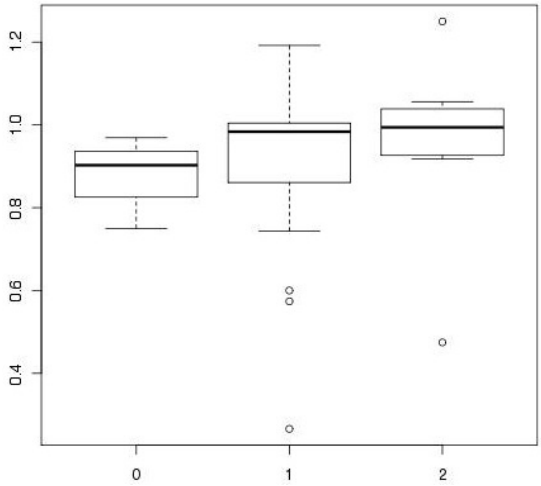


Figure 7. Boxplot of normalised sociability in German (outward interactions) and improvement by levels

(ii) Participants who did not show improvement had fewer general incoming interactions, but more German incoming interactions. The latter effect is even more prominent when framed in the context of the former. This finding may first seem counterintuitive (suggesting that more incoming German interactions are associated with poorer performance). However, if we remember the fact that for each participant, incoming interaction scores are dependent on the reports of *other*, it follows that those receiving more incoming interactions are at the same time enabling others to have more outgoing interactions (in other words, they are being 'used' by others for speaking German; *cf.* Fig. 8);

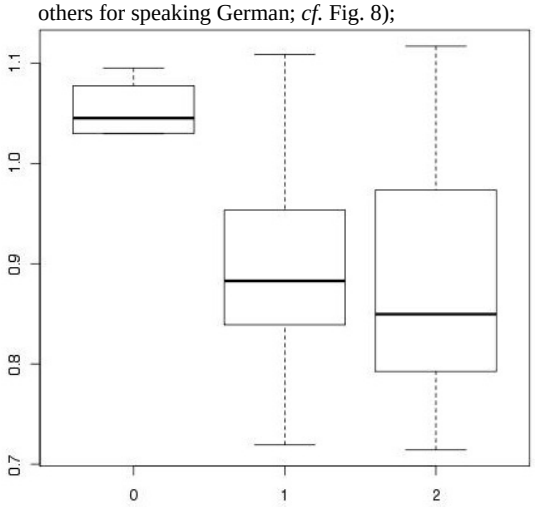


Figure 8. Boxplot of normalised popularity in German (incoming interactions) and improvement by levels

(iii) Neither incoming nor outgoing German interactions alone are strongly associated with homophily in performance. However, when both are considered, the frequency of interaction between participants is strongly associated with similarity in their performance;

(iv) There appeared to be no relationship between general interactions and performance;

(v) There was a clear negative relationship between performance and the number of interactions with participants with the same native language such that participants who showed no improvement in level interacted significantly more with those sharing their native language than did the participants who improved by two levels. This effect was observed both for the general and the German interactions:

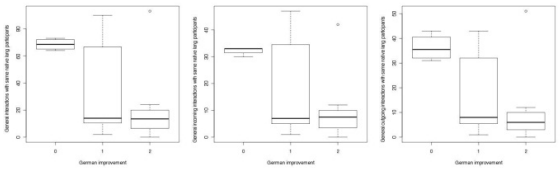


Figure 9. Boxplots of general interactions with same-native-language participants. Left: both incoming and outgoing, Centre: incoming, Right: outgoing

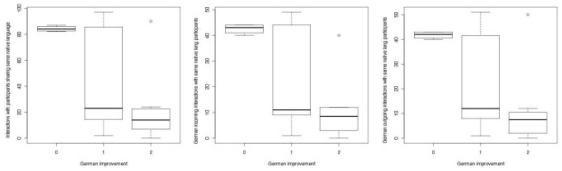


Figure 10. Boxplots of German interactions with same-native-language participants. Left: both incoming and outgoing, Centre: incoming, Right: outgoing

9 CONCLUSIONS

The results of social network analyses not only help understand social behaviour and determine the degree to which individual agents succeed in achieving their goals, but also provide useful indications for systems where non-human agents have to interact or teamwork with other artificial or human actors, machine learning and collective intelligence. The design of intelligent machines would benefit from seeing them as actors in a realistic social context, where the number, nature and influence of neighbours play an important part in the learning process. For instance, exposure thresholds and creativity ratios can constitute useful benchmarks for machines learning from and interacting with many other agents, while the finding that outgoing interactions in the acquired language are a better predictor of performance than incoming interactions support Swain's Output Hypothesis [32] and the emergent grammar theory [16] lying behind formalisms such as Fluid Construction Grammar [31], which is used in robotics.

REFERENCES:

[1] A. Bandura. *Social Learning Theory*. Prentice Hall, Englewood Cliffs, NJ (1977).

[2] X. Castelló, L. Loureiro-Porto, V.M. Eguíluz, M. San Miguel. The fate of bilingualism in a model of language competition. In: Takahashi, S., Sallach, D., Rouchier, J. (ds) *Advancing Social Simulation: The First World Congress*, pp. 83-94. Springer, Berlin (2007).

[3] A. Clauset, C.R. Shalizi, M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review 51*(4), 661-703 (2009).

[4] R. Dawkins. *The Selfish Gene*. Oxford University Press, Oxford (1976).

[5] J. Davies. Display, identity and the everyday: Self-presentation through digital image sharing. *Discourse, Studies in the Cultural Politics of Education 28*(4), 549-64 (2007).

[6] P. Benson. SLA after YouTube: New literacies and new language learning. Inv. talk, Univ. Warsaw (22 Oct 2010).

[7] G. de Tarde. *Les lois de l'imitation: étude sociologique*. Félix Alcan, Paris (1890).

[8] D.C. Dennett. Memes and the Exploitation of Imagination. *J Aesthetics and Art Criticism 48*(2), 127-135 (1990).

[9] A. Firth, J. Wagner. On discourse, communication, and (some) fundamental concepts in SLA research. *The Modern Language J* 81, 285-300 (1997).

[10] A. Firth, J. Wagner. Second/Foreign language learning as a social accomplishment: Elaborations on a 'reconceptualized' SLA. *The Modern Language J* 91, 800-19 (2007).

[11] M. Gladwell. Small Change: Why the Revolution Will Not Be Tweeted. *The New Yorker* (4 Oct 2010).

[12] Global Language Monitor. Death of Michael Jackson. http://www.languagemonitor.com/news/death-of-michael-jackson/ (29 Jun 2009).

[13] T. Gong, J.W. Minett, W.S.-Y. Wang. Exploring social structure effect on language evolution based on a computational model. *Connection Science 20*(2), 135-53 (2008).

[14] M. Granovetter. Threshold Models of Collective Behavior. *The Am J Sociology 83*(6), 1420-43 (1978).

[15] J. Henrich, R. Boyd. On modeling cognition and culture: Why cultural evolution does not require replication of representations. *J Cognition and Culture 2*(2), 87-112 (2002).

[16] P. Hopper. Emergent Grammar. *Procs Berkeley Linguistics Soc* 13, 139-57 (1987).

[17] R. Jakobson. *Essais de linguistique générale*. Éditions de Minuit, Paris (1963). [English version in same, *Selected Writings: Word and Language*, Vol. 2. Mouton, The Hague (1971).]

[18] M. Knobel, C. Lankshear. Online memes, affinities, and cultural production. In C. Lankshear, M. Knobel, C. Bigum, M. Peters (eds) *A New Literacies Sampler*, Vol. 29, pp. 199-227. Peter Lang, New York (2007).

[19] C. Kramsch (Ed.) *Language Acquisition and Language Socialisation: Ecological Perspectives*. Continuum, London (2002).

[20] J.P. Lantolf, S.L. Thorne. *Sociocultural Theory and the Sociogenesis of Second Language Development*. Oxford University Press, New York (2006).

[21] P.F. Lazarsfeld, R.K. Merton. Friendship as a Social Process: A Substantive and Methodological Analysis. In M. Berger, T. Abel, C.H. Page (eds) *Freedom and Control in Modern Society*, pp.18-66. Van Nostrand, New York (1954).

[22] J. Leskovec, J., L. Backstrom, J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining (KDD) (2009).

[23] M. McPherson, L. Smith-Lovin, J.M. Cook. Birds of a Feather: Homophily in Social Networks. *Ann Rev Sociol* 27, 415-44 (2001).

[24] G. Merchant. Teenagers in cyberspace: an investigation of language use and language change in Internet chatrooms. *J Research in Reading 24*(3), 293-306.

[25] M.E.J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics 46*(5), 323-51 (2005).

[26] J. O'Dell. Does Online Buzz Mean Better TV Ratings? http://mashable.com/2010/06/24/neilsen-vs-social-media/ (24 Jun 2010).

[27] M.B. Paradowski, Ł. Jonak: Milgram's legacy and lipstick on a pig – how to track linguistic epidemics in online social networks. paper pres. at Intl School Conf on Network Sci, Central European Univ/Hungarian Acad Sciences, Budapest (10 Jun 2011).

[28] S. Pekarek Doehler. Conceptual changes and methodological challenges: on language and learning from a conversation analytic perspective on SLA. In P. Seedhouse, S. Walsh, Ch. Jenks (eds) *Conceptualising Learning in Applied Linguistics*, pp. 105-127. Palgrave Macmillan, Basingstoke (2010).

[29] Project for Excellence in Journalism. New media, old media. How Blogs and Social Media Agendas Relate and Differ from Traditional Press. http://www.journalism.org/analysis_report/new_media_old_media (23 May 2010).

[30] D. Sperber. An objection to the memetic approach to culture. In R. Aunger (Ed.) *Darwinizing Culture: The status of memetics as a science*, pp. 163-173. Oxford University Press, Oxford (2000).

[31] L. Steels, J. De Beule. A (very) Brief Introduction to Fluid Construction Grammar. *Procs 3rd Workshop on Scalable Natural Language Understanding*, pp. 73-80. Assoc for Computational Linguistics, New York (8 Jun 2006).

[32] M. Swain. Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass, C. Madden (eds) *Input in Second Language Acquisition*, pp. 235-56. Newbury House, New York (1985).

[33] M. Tamariz, T. Gong, G. Jaeger. Investigating the effects of prestige on the diffusion of linguistic variants. In L. Carlson, C. Hoelscher, T.F. Shipley (eds) *Expanding the Space of Cognitive Science. Procs 33rd Ann Meeting Cognitive Science Soc*, pp. 1491-6. Cognitive Science Society, Austin, TX (2011).

[34] T.W. Valente. *Network Models of the Diffusion of Innovations*. Hampton Press, Cresskill, NJ (1995).

[35] Th. Vander Wal. Folksonomy coinage and definition. http://vanderwal.net/folksonomy.html (2 Feb 2007).